

Нейроинформатика, её приложения и анализ данных

XXV Всероссийский семинар

**Красноярск
2017**

Министерство образования и науки РФ
Сибирское отделение РАН
Российская ассоциация нейроинформатики
Институт вычислительного моделирования СО РАН

НЕЙРОИНФОРМАТИКА, ЕЁ ПРИЛОЖЕНИЯ И АНАЛИЗ ДАННЫХ

МАТЕРИАЛЫ
ДВАДЦАТЬ ПЯТОГО ВСЕРОССИЙСКОГО СЕМИНАРА
29 сентября – 1 октября 2017 года

Красноярск 2017

УДК 007.52 (042.3)

Н 43

Нейроинформатика, её приложения и анализ данных: Материалы XXIV Всероссийского семинара, 29 сентября – 1 октября 2017 года / Под ред. А.Н.Горбаня, отв. за вып. М.Ю. Сенашова; – Красноярск: Институт вычислительного моделирования СО РАН, 2017. – 125 с.

В сборнике представлены материалы XXV Всероссийского семинара «Нейроинформатика, её приложения и анализ данных», проходившей в городе Красноярске 29 сентября – 1 октября 2017 года

Основной задачей конференции является всесторонний и высоко квалифицированный обмен новейшими достижениями в различных областях нейроинформатики (как теории нейросетей, так и в области различных практических приложений), а также в области нелинейного статистического анализа многомерных данных, обладающих нетривиальными структурами.

Большое внимание уделено анализу областей применимости и точности методов обработки многомерных данных, анализу устойчивости различных новых (нелинейных) методов кластеризации, разбор большого числа конкретных случаев, иллюстрирующих эти проблемы и достижения.

Материалы предназначены для научных работников, преподавателей, студентов и аспирантов соответствующих специальностей.

Редакционная коллегия:

Горбань Александр Николаевич – ответственный редактор

Сенашова Мария Юрьевна – ответственный за выпуск

© ИВМ СО РАН, 2017

© Коллектив авторов, 2017

ISBN 978-5-906455-11-6



9 785906 455116

УСКОРЕННЫЕ МЕТОДЫ ПЕРВОГО ПОРЯДКА ОБУЧЕНИЯ СЕТЕЙ РАДИАЛЬНЫХ БАЗИСНЫХ ФУНКЦИЙ¹

М.М. Алкезуини, В.И. Горбаченко

Пензенский государственный университет,
факультет вычислительной техники,
кафедра "Компьютерные технологии", gorvi@mail.ru

Сеть радиальных базисных функций (РБФ-сеть) — двухслойная сеть, первый слой которой составляют радиальные базисные функции (РБФ), второй — линейный сумматор [1, 2]. Выход РБФ-сети описывается выражением

$$u(\mathbf{x}) = \sum_{i=1}^{n_{RBF}} w_i \varphi_i(\mathbf{x}), \quad (1)$$

где $\mathbf{x}$ — входной вектор, w_i — вес i -й РБФ, $\varphi_i(\mathbf{x})$ — значение i -й РБФ для входного вектора $\mathbf{x}$, n_{RBF} — количество РБФ.

В качестве РБФ используются различные радиальные базисные функции. В данной работе применяется гауссиан

$$\varphi_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_i\|^2}{2a_i^2}\right), \quad (2)$$

где $\mathbf{c}$ — вектор координат центра i -ой RBF, a — ширина (параметр формы) RBF, $\|\mathbf{x} - \mathbf{c}_i\|$ — евклидова норма (расстояние) между входным вектором и центром RBF.

Радиальные базисные функции широко используются при решении задач аппроксимации функций [3], в частности, аппроксимации "рассеянных" данных, когда узлы интерполяции расположены произвольным образом, а не на некоторой сетке. Такие задачи возникают, например, при моделировании рельефа и реконструкции поверхности. РБФ применяются в бессеточных методах решения краевых задач, описываемых дифференциальными уравнениями в частных производных [4]. Недостатком применения РБФ является трудно формализуемый подбор параметров РБФ. Этот недостаток устраняется при использовании РБФ-сетей. При использовании РБФ-сетей решение формируется в процессе обучения сети с настройкой весов и параметров РБФ. В частности, разработаны методы решения краевых задач на РБФ-сетях [5–7].

Рассмотрим обучение РБФ-сети при решении задач аппроксимации функций, когда необходимо аппроксимировать функцию, заданную значениями в произвольно расположенных узлах интерполяции. В этом случае входные векторы при обучении сети представляют собой координаты пробных точек, а целевые значения — значения функции в пробных точках. После обучения сеть при подаче координат произвольной точки области определения функции выдает значение функции в этой точке.

Обучение РБФ-сетей имеет свои особенности. Наличие только двух слоев упрощает обучение сети. Но в процессе обучения РБФ-сети необходимо настраивать две группы параметров сети: веса, которые входят линейно в выходной сигнал сети (1), и параметры РБФ (в случае гауссиана (2) — центры и ширина), которые входят нелинейно в выходной сигнал. Открытым является вопрос выбора количества РБФ и начальных значений параметров РБФ. Известна зависимость между количеством РБФ n_{RBF} для задач аппроксимации, и количеством узлов интерполяции N : $n_{RBF} \propto N^{1/3}$, где $\propto$ означает пропорциональность [8]. Однако эта оценка очень завышена и количество РБФ требуется подбирать. Если в процессе обучения сети настраиваются параметры РБФ, то их начальные значения целесообразно выбирать случайным образом.

Обучение РБФ-сети представляет собой минимизацию некоторого функционала ошибки (функции потерь в терминах машинного обучения). Будем использовать функционал ошибки, представляющий собой сумму квадратов ошибок в пробных точках

¹ Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 16-08-00906

$$I = \frac{1}{2} \sum_{j=1}^N (u(\mathbf{p}_j) - t_j)^2, \quad (3)$$

где N — количество пробных точек, $\mathbf{p}_j$ — координаты j -й пробной точки, t_j — целевое значение в j -й пробной точке, множитель $1/2$ введен для упрощения вычислений.

Для обучения сетей РБФ используются, в основном, градиентные методы. Среди градиентных методов выделяют методы первого порядка, использующие для своей реализации первые производные минимизируемой функции, то есть градиент функции, и методы второго порядка, использующие вторые производные — матрицу Гессе. Все градиентные алгоритмы позволяют найти только локальный минимум функции. Методы первого порядка просты в реализации, но работают медленно. Методы второго порядка выполняются за гораздо меньшее число итераций, но сложны и ресурсоемки, то есть одна итерация занимает много времени и требует много памяти. Для обучения сетей РБФ используют, в основном, методы первого порядка. Известные методы второго порядка (например, предложенный в [7] быстрый алгоритм метода доверительных областей) весьма сложны.

В настоящее время для глубокого обучения сетей, содержащих большое количество слоев и обучающихся на больших объемах данных, усилился интерес к несложным ускоренным градиентным методам первого порядка [9]. Такие методы не применялись для обучения сетей РБФ. Целью настоящей работы является экспериментальное исследование применимости ускоренных градиентных методов первого порядка для обучения сетей РБФ при решении задач аппроксимации.

Рассмотрим алгоритм градиентного спуска в общем виде. Если I — функционал ошибки (3), а Θ — вектор одного или всех параметров сети (в нашем случае это векторы весов, центров и ширины РБФ), то настройка вектора Θ на k -ой итерации градиентного спуска описывается следующим образом [2]

$$\Theta^{(k+1)} = \Theta^{(k)} + \Delta\Theta^{(k+1)}, \quad (4)$$

где $\Delta\Theta^{(k+1)} = -\eta \mathbf{g}_\Theta(\Theta^{(k)})$ — поправка вектора Θ , η — подбираемый числовой коэффициент (скорость обучения), $\mathbf{g}_\Theta(\Theta^{(k)})$ — вектор градиента функционала I по значению параметра Θ на итерации k .

Процесс вычислений по (4) продолжается до малого значения функционала ошибки.

В алгоритме градиентного спуска с импульсом (в русской литературе принято использовать термин "момент" [2], принятый в англоязычной литературе термин "momentum — импульс" по аналогии с импульсом в механике точнее отражает суть метода) [2, 9] поправка к вектору параметров описывается следующим образом

$$\Delta\Theta^{(k+1)} = \alpha \Delta\Theta^{(k)} - \eta \mathbf{g}_\Theta(\Theta^{(k)}), \quad (5)$$

где η — скорость обучения, α — коэффициент момента, принимающий значения в интервале $[0, 1]$.

Выражение (5) содержат слагаемые, зависящие от градиента, и не зависящие от градиента. Причем чем больше значение коэффициента α , тем сильнее на настройку весов сказывается слагаемое, не зависящее от градиента. Это влияние существенно возрастает на плоских участках целевой функции и вблизи локальных минимумов. В этих областях слагаемое, не зависящее от градиента, начинает доминировать в (5), что приводит к выходу из этой области. Можно сказать, что изменение весов происходит "по инерции" в направлении вектора поправки предыдущего шага обучения.

В методе ускоренного градиента Нестерова (NAG — Nesterov Accelerated Gradient) [9–10] поправка вектора параметров описывается следующим образом

$$\Delta\Theta^{(k+1)} = \alpha \Delta\Theta^{(k)} - \eta \mathbf{g}_\Theta(\Theta^{(k)} + \alpha \Delta\Theta^{(k)}).$$

Метод NAG отличается от градиентного спуска с импульсом более точным вычислением вектора поправки.

При обучении глубоких сетей получили распространение алгоритмы с адаптивной скоростью обучения [9]. В этих алгоритмах используется различная скорость обучения для различных компонентов вектора параметров, так как разные компоненты

могут оказывать различное влияние на обучение сети. В частности, эффективным и практичным методом является RMSProp (root mean square propagation) и его сочетание с ускоренным градиентом Нестерова [9], k -я итерация метода включает следующие вычисления:

$$\mathbf{g} = \mathbf{g}_\Theta(\Theta^{(k)} + \alpha\Delta\Theta^{(k)}), \quad \mathbf{r}^{(k+1)} = \rho\mathbf{r}^{(k)} + (1 - \rho)\mathbf{g}\bullet\mathbf{g},$$

$$\Delta\Theta^{(k+1)} = \alpha\Delta\Theta^{(k)} - \frac{\varepsilon}{\sqrt{\mathbf{r}^{(k+1)}}}\bullet\mathbf{g}, \quad \Theta^{(k+1)} = \Theta^{(k)} + \Delta\Theta^{(k+1)},$$

где $\mathbf{r}^{(0)} = \mathbf{0}$, $\bullet$ — операция поэлементного умножения (произведение Адамара), $\sqrt{\mathbf{r}^{(k+1)}}$ вычисляется поэлементно, $\rho, \alpha, \varepsilon$ — подбираемые коэффициенты.

При обучении сети РБФ можно реализовать одновременную и последовательную стратегию настройки векторов весов, центров и ширины на каждом цикле обучения. При одновременной настройке в каждой итерации обучения при значениях векторов параметров, полученных на предыдущей итерации, находятся векторы градиентов функционала ошибки по каждому параметру, затем вычисляются поправки параметров и вычисляются обновленные векторы параметров. При последовательной настройке сначала вычисляется градиент по весам (настройка весов оказывает наибольшее влияние на обучение сети) и вычисляется обновленный вектор весов. Затем с использованием обновленного вектора весов вычисляется градиент по центрам и вычисляется обновленный вектор центров. Наконец, с использованием обновленных векторов весов и центров вычисляется градиент по ширине и обновляется вектор ширины. Так как при последовательной настройке каждый параметр настраивается при обновленной части других параметров, можно предположить, что последовательная настройка может обеспечить сокращение числа итераций обучения.

Экспериментальное исследование рассмотренных методов проводилось на примере аппроксимации функции $z = x^2 + y^2$ в области $(x = -3 \dots +3, y = -3 \dots +3)$. Количество узлов интерполяции равно 100. Узлы интерполяции располагались случайным образом в области (рис. 1). Количество нейронов равно 16. В начальном состоянии центры РБФ располагались на сетке (рис. 2).

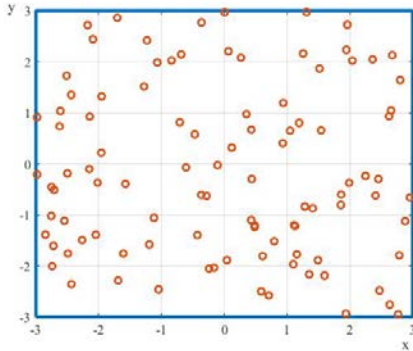


Рис. 1. Расположение узлов интерполяции

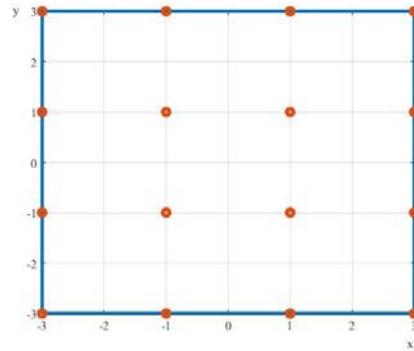


Рис. 2. Начальное расположение центров РБФ

Веса инициализировались случайными числами, равномерно распределенными от 0 до 0,001. Начальная ширина всех РБФ равно 3,0. Итерационный процесс обучения продолжался до выполнения условия для средней квадратической ошибки (Mean Squared Error) аппроксимации

$$\frac{1}{N} \sum_{j=1}^N (u(\mathbf{p}_j) - t_j)^2 \leq 0,01,$$

где N — количество пробных точек, $\mathbf{p}_j$ — координаты j -й пробной точки, t_j — целевое значение в j -й пробной точке.

Компоненты градиента функционала по весу, центрам и ширине вычислялись аналитически (номера итераций опущены)

$$\frac{\partial I}{\partial w_i} = \sum_{j=1}^N (u(\mathbf{p}_j) - t_j) \varphi_i(\mathbf{p}_j), \quad \frac{\partial I}{\partial c_{ix}} = w_i \sum_{j=1}^N (u(\mathbf{p}_j) - t_j) \varphi_i(\mathbf{p}_j) \frac{p_{jx} - c_{ix}}{a_i^2},$$

$$\frac{\partial I}{\partial c_{iy}} = w_i \sum_{j=1}^N (u(\mathbf{p}_j) - t_j) \varphi_i(\mathbf{p}_j) \frac{p_{jy} - c_{iy}}{a_i^2}, \quad \frac{\partial I}{\partial a_i} = w_i \sum_{j=1}^N (u(\mathbf{p}_j) - t_j) \varphi_i(\mathbf{p}_j) \frac{\|\mathbf{p}_j - \mathbf{c}\|^2}{a_i^3},$$

где c_{ix} и c_{iy} — координаты по осям x и y центра i -ой РБФ, p_{jx} и p_{jy} — координаты по осям x и y пробной точки $\mathbf{p}_j$.

Эксперименты проводились в системе MATLAB на компьютере со следующими характеристиками: процессор Intel Core i5-2500K, 3,30 GHz, ОЗУ 8,0 GB. Результаты экспериментов представлены в таблице. Так как число итераций и время решения зависят от случайных начальных значений весов, то для каждого метода проводилось 10 экспериментов и в таблице представлены полученные диапазоны числа итераций и времени решения. Индексы у коэффициентов означают: 1 — коэффициент для весов, 2 — для центров, 3 — для ширины. Значения коэффициентов подбирались экспериментально.

Таблица

Результаты экспериментов

Метод	Стратегия обучения	Параметры	Число итераций	Время решения, С
Градиентный спуск	Одновременная настройка	$\eta_1 = 0,0015,$ $\eta_2 = 0,0010,$ $\eta_3 = 0,0005$	45000–70000	2500–5000
Градиентный спуск	Последовательная настройка	$\eta_1 = 0,0500,$ $\eta_2 = 0,0010,$ $\eta_3 = 0,0005$	35000–50000	2000–3500
Градиентный спуск с импульсом	Одновременная настройка	$\eta_1 = 0,00700, \alpha_1 = 0,9,$ $\eta_2 = 0,00002, \alpha_2 = 0,9,$ $\eta_3 = 0,00020, \alpha_3 = 0,9$	1500–2100	50–70
Градиентный спуск с импульсом	Последовательная настройка	$\eta_1 = 0,00700, \alpha_1 = 0,9,$ $\eta_2 = 0,00002, \alpha_2 = 0,9,$ $\eta_3 = 0,00020, \alpha_3 = 0,9$	200–9000	9–360
NAG	Последовательная настройка	$\eta_1 = 0,0050, \alpha_1 = 0,9$ $\eta_2 = 0,0020, \alpha_2 = 0,5$ $\eta_3 = 0,0010, \alpha_3 = 0,3$	270–470	14–24
RMSProp+NAG	Последовательная настройка	$\eta_1 = 0,0010,$ $\eta_2 = 0,0020,$ $\eta_3 = 0,0010,$ $\alpha_1 = 0,9, \rho_1 = 0,9$ $\alpha_2 = 0,5, \rho_2 = 0,9$ $\alpha_3 = 0,1, \rho_3 = 0,9$	5700–13400	245–600

Результаты экспериментов показали явное преимущество ускоренных методов по сравнению с классическим методом градиентного спуска. Последовательная настройка параметров также дает существенный эффект. Градиентный спуск с импульсом очень чувствителен к случайным начальным значениям весов и параметрам обучения. Процесс изменения средней квадратической погрешности носит колебательный характер и неустойчив. Наиболее быстрым и устойчивым является метод NAG. Метод наименее чувствителен к начальным значениям весов и параметрам обучения. Процесс изменения средней квадратической погрешности носит гладкий характер и устойчив (рис. 3). Причем экспериментально обнаружена необходимость использования различных значений коэффициентов α для различных параметров сети. Метод RMSProp+NAG

устойчив к изменению начальных значений весов и параметров обучения, Процесс изменения средней квадратической погрешности носит гладкий характер и устойчив. Но метод проигрывает NAG по числу итераций.

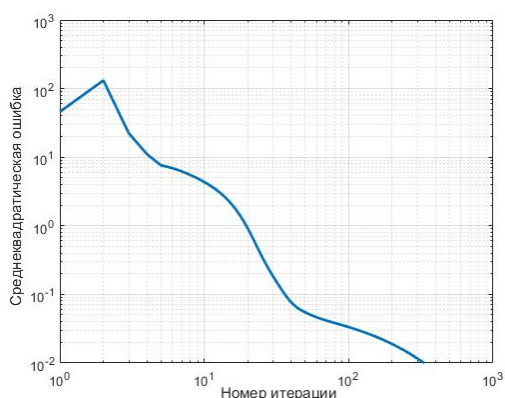


Рис. 3. Зависимость средней квадратической ошибки от номера итерации в методе NAG

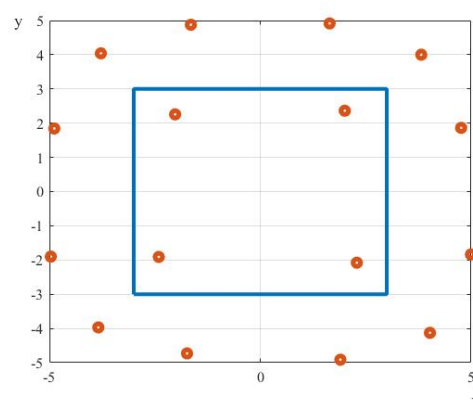


Рис. 4. Конечное положение центров РБФ

Эксперименты подтвердили важность настройки не только весов, но и параметров РБФ. Конечное положение центров РБФ, полученное в результате обучения сети (рис. 4) радикально отличается от начального положения (рис. 3), причем после обучения сети центры вышли за пределы области решения.

Таким образом, для обучения сетей радиальных базисных функций можно рекомендовать ускоренный метод градиента Нестерова с последовательной настройкой параметров сети.

Список литературы

1. Хайкин С. Нейронные сети: полный курс. — М.: Вильямс, 2006. — 1104с.
2. Осовский С. Нейронные сети для обработки информации. — М.: Горячая линия-Телеком, 2016. — 448 с.
3. Fasshauer G. F. Meshfree Approximation Methods with MATLAB. — World Scientific Publishing Company, 2007. — 520 p.
4. Chen W., Fu Z.-J., Chen C. S. Recent Advances in Radial Basis Function Collocation Methods. — Springer, 2014. — 90 p.
5. Yadav N., Yadav M., Kumar M. An Introduction to Neural Network Methods for Differential Equations. — Springer, 2015. — 115 p.
6. Васильев А. Н., Тархов Д. А. Принципы и техника нейросетевого моделирования. — СПб.: Нестор-История, 2014. — 218 с.
7. Горбаченко В. И., Жуков М. В. Решение краевых задач математической физики с помощью сетей радиальных базисных функций // Журнал вычислительной математики и математической физики. — 2017, том 57. — № 1. — С. 133–143.
8. Niyogi P., Girosi F. On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions // Neural Computation. — 1996. — Vol. 8, Issue 4. — P. 819–842.
9. Goodfellow I., Bengio Y., Courville A. Deep Learning. — MIT Press, 2016. — 800 p.
10. On the importance of initialization and momentum in deep learning / I. Sutskever, J. Martens, G. Dahl, G. Hinton // Proceedings of the 30th International Conference on International Conference on Machine Learning Volume. — 2013. — Vol. 28. — P. 1139–1147.

ОПРЕДЕЛЕНИЕ ПОКАЗАТЕЛЕЙ УСТОЙЧИВОГО ФУНКЦИОНИРОВАНИЯ НЕЙРОПОДОБНЫХ СИСТЕМ

С.С. Анцыферов, К.Н. Фазилова, К.Е. Русанов

Московский технологический университет,
Физико-технологический институт, каф. метрологии и стандартизации,
fazilova@mirea.ru

Введение

Нейроподобные системы обладают рядом свойств, таких как: устойчивость, целенность, надежность, адаптивность, быстродействие, самоорганизованность. В последнее время ведется работа как по изучению показателей свойств НС, так и по их стандартизации [1-12].

При описании показателей свойств нейроподобных систем используют такие модели нейронных сетей, как модели Хопфилда, «Машина Больцмана», нейробиологические модели Палма, сложная нейронная модель Фукусима. В настоящее время наиболее распространенной является модель Хопфилда.

Согласно данной модели, при описании характера функционирования нейронной сети используются переменные, описывающие состояние нейронов (S), и переменные, описывающие связи между нейронами (J). Обучение нейроподобных систем связано с возникновением положительных обратных связей, которые устанавливаются между теми нейронами, которые при решении одной задачи, оказываются в одинаковом (например, возбужденном) состоянии. В модели это должно соответствовать увеличению веса связи между i -м и j -м нейронами J_{ij} . Заметим, что при обучении нейроподобной системы по правилу Хебба, помимо «настоящих» аттракторов, возможно формирование аттракторов не совпадающих ни с одним из ключевых образов, т.е. ложных образов. Принятыми показателями устойчивости НС, согласно модели Хопфилда, являются такие параметры как

$A = M/N$, где M – число ключевых образов, которые могут быть сохранены,

N – число нейронов сети;

$R = \frac{1}{N} \sum_{i=1}^N \varphi_i \varphi_j$ – коэффициент корреляции между образами (чем больше

корреляция нового образа с одним из уже записанных, тем сложнее нейронной сети запомнить, распознать его и сохранить при этом высокую устойчивость), где φ_i , φ_j – сформированные образы.

H^* – энтропия, характеризующая степень размытости ключевых образов системы,

λ – сила связи между нейронами (нейроподобными элементами).

Вместе с тем, модель Хопфилда является отражением сенсорного характера взаимодействия нейроподобных элементов, что существенно ограничивает процесс целенаправленной трансформации структуры системы при изменении содержания и интенсивности входных информационных потоков. Поэтому представляется целесообразным для обеспечения способности систем к самоорганизации и адаптации рассматривать нейроподобный элемент как активный, обладающий определенной базой знаний и способный принимать самостоятельные решения. При этом полагается, что каждый элемент выполняет свою независимую часть решения задачи, его знание ограничено текущими задачами, он не владеет информацией обо всех файлах

одновременно (знание ограничено только обрабатываемым файлом) или физических границах предметной области.

Активные элементы (АЭ), как и в модели Хопфилда, могут попарно взаимодействовать между собой, множество элементов структурировано, каждый элемент, решая свою часть общей задачи, координирует общий ход решения с другими элементами. В результате, окончательное решение можно рассматривать как коллективное скоординированное (согласованное), не являющееся суммой отдельных решений, что служит проявлением такого свойства НС как целостность (эмерджентность). В настоящее время отсутствуют модели функционирования систем с активными элементами, позволяющие устанавливать показатели устойчиво-эффективного функционирования.

Цель данной работы – разработка модели функционирования нейроподобной системы с активными элементами, обеспечивающей возможность установления показателей устойчивого функционирования системы. Однако, прежде следует определить модель функционирования отдельного АЭ.

Модель функционирования активного элемента

При построении модели функционирования АЭ будем исходить из принципов продукционного программирования, согласно которым программа задается совокупностью правил без явного указания последовательности их выполнения. Правила, являющиеся по-существу формой предоставления знаний, содержат либо условие и действие, которое должно быть выполнено в случае истинности этого условия, либо условие и совокупность других условий, достаточных для истинности этого условия. В соответствии с данными принципами, АЭ должен обладать рабочей памятью, в которой содержится описание текущего состояния решаемой части задачи. Это описание служит образцом, с которым сопоставляется условная часть продукции с целью выбора соответствующих действий при решении задачи. Если условие некоторого правила соответствует содержимому рабочей памяти, то может выполняться действие, связанное с этим условием. Действия продукционных правил приводят к изменению содержания рабочей памяти. Таким образом, алгоритм функционирования АЭ состоит в следующем:

- * рабочая память инициализируется начальным описанием задачи. Текущее состояние решения задачи представляется набором образцов в рабочей памяти;

- * образцы сопоставляются с условиями продукционных правил, что порождает подмножество правил вывода (конфликтное подмножество), условия которых согласованы с образцами в рабочей памяти;

- * выбирается и активируется одна из продукций конфликтного множества (разрешение конфликта). Активизация правила означает выполнение его действия, в результате которого изменяется содержание рабочей памяти.

После того, как выбранное правило сработало, цикл управления повторяется для модифицированной рабочей памяти. Процесс заканчивается, если содержимое рабочей памяти не соответствует никаким условиям. В результате функционирования АЭ приобретает циклический характер (рис. 1):

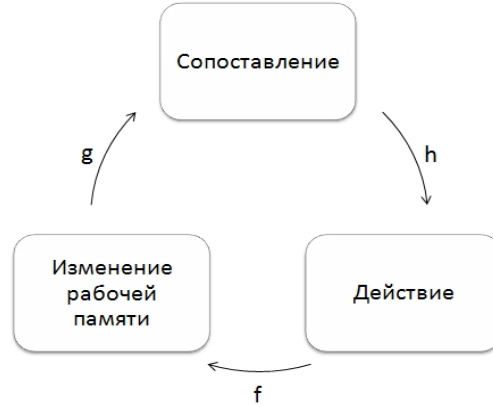


Рис. 1. Функционирование активного элемента

Эффективность функционирования АЭ определяется мерой достоверности «заложеной» в него базы знаний. Естественно предполагать, что данная мера является вероятностной и формально может быть определена согласно рекуррентному правилу Байеса:

$$p(\hat{\mu}_{jk}) = \int_M \tilde{p}(\hat{\mu}_{jk-1}) p(\hat{\mu}_{jk} | \tilde{\mu}_{jk-1}) d\tilde{\mu}_{jk-1};$$

$$\hat{\mu}_{jk} \equiv \tilde{p}(\hat{\mu}_{jk}) = \frac{p(\hat{\mu}_{jk}) p(I_{jk} | \hat{\mu}_{jk})}{\int_M p(\hat{\mu}_{jk}) p(I_{jk} | \hat{\mu}_{jk}) d\hat{\mu}_{jk}}.$$

Действия каждого элемента в рамках байесовской модели будем аппроксимировать некоторым набором детерминированных функций, позволяющих построить (рис. 2) фазовый портрет функционирования АЭ: $a_1 = h(a_3)$, $a_2 = f(a_1)$, $a_3 = g(a_2)$, где a_1 – действие, a_2 – результат действия (изменение рабочей памяти), a_3 – оценка результата (сопоставление): $h(a) = k(1 - e^{-ka})$; $f(a) = ke^{-ka}$; $g(a) = \log_x a$.

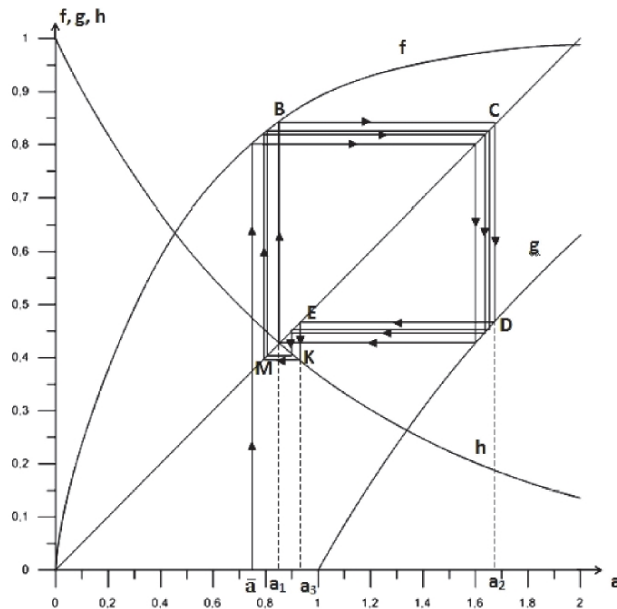


Рис. 2. Фазовый портрет функционирования активного элемента

Используя предложенный фазовый портрет, при заданных значениях параметров функций ($k_1; k_2; x$ – основание логарифма) возможно установление пределов допустимых соотношений между базой знаний и объемом рабочей памяти.

При выходе за допустимый диапазон происходит распад предельного цикла, а следовательно и возможная потеря устойчивости.

Модель функционирования нейрореподобной системы

Адекватной моделью описания состояния нейрореподобной системы с активными элементами может служить энтропия, в частности энтропия их попарных взаимодействий:

$$H = H_1 + H_2 = \left[- \sum_i P_i \log P_i \right] + \left[- \sum_i \sum_j P_{ij} \log P_{ij} \right] -$$

где P_i - вероятность меры информативности i -го элемента;

P_{ij} - вероятность попарного взаимодействия.

Дифференциальное уравнение, описывающее динамику системы может иметь, согласно [7], следующий вид

$$\dot{H} = \frac{dH}{dt} = KH,$$

где $K = \Delta I - \Delta JH$,

ΔI – параметр, характеризующий управление процессом формирования структуры, показатель функциональной организации системы.

$$\begin{aligned} \dot{H} &= \Delta IH - \Delta JH^2 = -\Delta J \left(H^2 - \frac{\Delta I}{\Delta J} H \right) = \\ &= -\Delta J \left[H^2 - \frac{\Delta I}{\Delta J} H + \left(\frac{1}{2} \frac{\Delta I}{\Delta J} \right)^2 - \frac{\Delta I^2}{4\Delta J^2} \right] = -\Delta J \left[\left(H - \frac{\Delta I}{2\Delta J} \right)^2 - \frac{\Delta I^2}{4\Delta J^2} \right], \end{aligned}$$

где ΔI – приращение интенсивности информационного потока (увеличение семантической, т.е. содержательной составляющей);

ΔJ – приращение интенсивности обработки информационного потока (увеличение меры достоверности базы знаний системы).

Корни данного уравнения: $H_1 = 0$, $H_2 = \frac{\Delta I}{\Delta J}$.

H_1 и H_2 – точки устойчивости (рис. 3).

При $\Delta I = 0$ и $\Delta J = 0$, ΔH и $\dot{H}$ равны нулю и система находится в устойчивом состоянии (точка H_1).

Возрастание интенсивности информационного потока ($\Delta I > 0$) приводит к изменению энтропии системы ($\Delta H > 0$) и адекватная реакция системы состоит в усилении действия ($\Delta J > 0$). В результате «включается» механизм положительной обратной связи ($\Delta H > 0, \dot{H} > 0$) и система переходит в новое устойчивое состояние (точка H_2), когда $\Delta I > 0$ и $\Delta J > 0$. В этом состоянии действует механизм отрицательной обратной связи ($\Delta H > 0, \dot{H} < 0$; $\Delta H < 0, \dot{H} > 0$). Это означает, что при всех флуктуациях H ($\Delta H < 0$, $\Delta H > 0$) система будет возвращаться в устойчивое состояние, т.е. в точку H_2 . Функционирование системы в точке H_2 будет носить устойчивый характер до тех пор пока $\Delta I > 0$ и $\Delta J > 0$. При $\Delta I < 0$ и $\Delta J < 0$ начинает действовать механизм положительной обратной связи (рис. 4) и в случае отсутствия противодействия такому

механизму система будет «уходить» из точки H_2 в точку H_1 при $\Delta H < 0$. Это может сопровождаться усилением связей (λ) между отдельными элементами. В предельном (гипотетическом) случае при $\lambda \rightarrow \infty$ система деградирует, т.е. перестает адекватно реагировать на изменение информационного потока. Если же $\Delta H > 0$, то система удаляется из области устойчивого функционирования в сторону возрастания энтропии, при этом связи между элементами ослабляются и в пределе, когда $\lambda = 0$ система распадается на N независимых элементов.

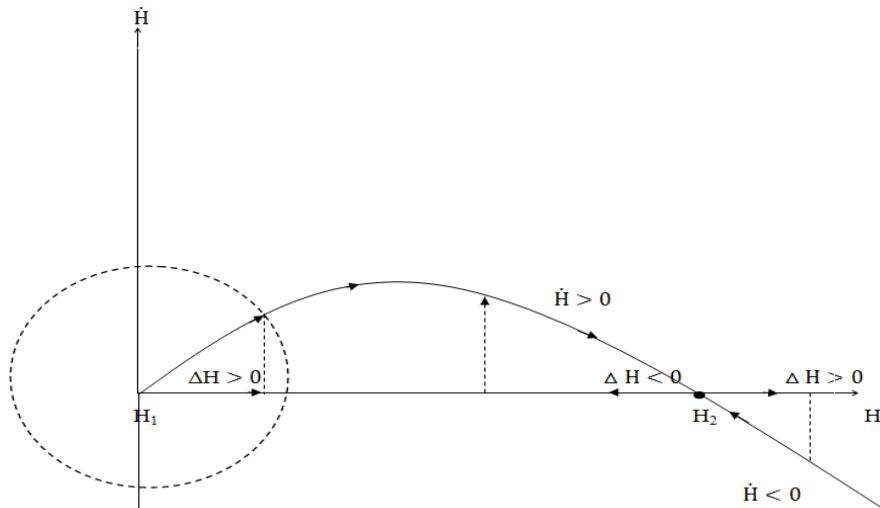


Рис. 3. Устойчивое функционирование НС

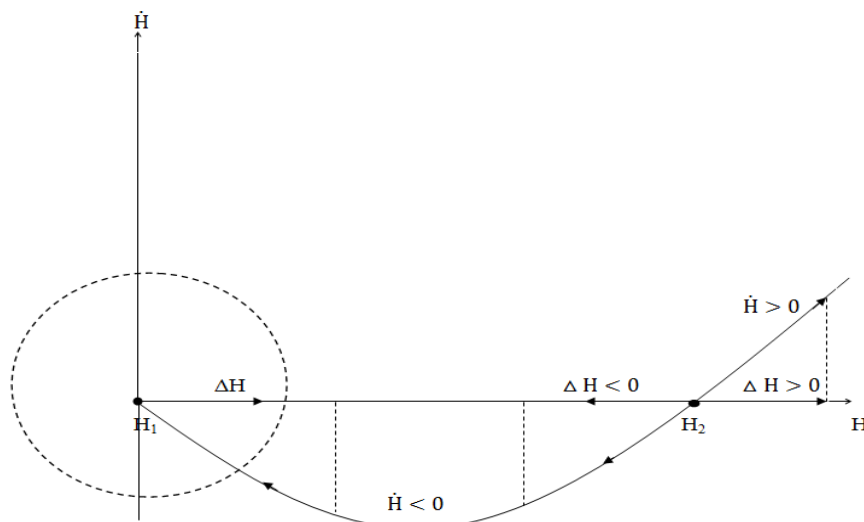


Рис. 4. Неустойчивое функционирование НС

Выводы

Предложенная модель функционирования нейроподобных систем с активными элементами позволяет определить основные показатели устойчивости:

* $A = M/N$, где M – число ключевых образов, которые могут быть сохранены,
 N – число нейронов сети;

- * H^* – энтропия, характеризующая степень размытости ключевых образов системы;
- * функции $h(a)$; $f(a)$; $g(a)$, отражающие характер функционирования АЭ;
- * соотношение приращений содержательной части внешнего и внутреннего информационных потоков $(\Delta I; \Delta J)$;
- * энтропия системы H ;
- * параметр λ характеризующий силу взаимодействия между нейроподобными элементами;
- * $\hat{\mu}_{jk}$ – мера информативности.

Список литературы

1. Анцыферов С.С. Формирование спектра тепловых изображений объектов и распознавание их образов // Оптический журнал. — 1999. — Т.66. — № 12. — С.44-48.
2. Анцыферов С.С. Формирование спектра тепловых изображений объектов и распознавание их образов // J. Opt. Technol. — 1999. — V.66. — № 12. — P.1047-1049.
3. Анцыферов С.С., Евтихий Н.Н., Голубь Б.И. Адаптация информационно-распознающих биомедицинских систем // Биомедицинская радиоэлектроника. — 2001. — № 1. — С.5-9.
4. Анцыферов С.С. Метрология виртуальных систем // Измерительная техника. — 2003. — № 5. — С.17-21.
5. Анцыферов С.С., Евтихий Н.Н. Адаптивные системы распознавания образов пространственно-временных полей // Искусственный интеллект. — 2004. — № 3. — С.405-416.
6. Анцыферов С.С., Евтихий Н.Н. Адаптивная обработка информации пространственно-временных изотропных полей // Оптический журнал. — 2006. — Т.3. — № 10. — С.52-57.
7. Анцыферов С.С. Общие принципы построения и закономерности функционирования интеллектуальных систем // Искусственный интеллект. — 2011. — № 3. — С.6-15.
8. Анцыферов С.С. Вопросы метрологического обеспечения интеллектуальных систем // Мир измерений. — 2012. — № 5. — С.46-51.
9. Анцыферов С.С. Метрологические основы формирования информационного пространства образов в интеллектуальных системах обработки // Искусственный интеллект. — 2012. — № 4(58). — С.283-291.
10. Анцыферов С.С., Русанов К.Е. Быстродействие интеллектуальных систем // Искусственный интеллект. — 2013. — № 4(62). — С.259-265.
11. Анцыферов С.С. Стандартизация показателей качества продукции когнитивных технологий // Наукоемкие технологии. — 2014. — Т.15. — № 7. — С.7-13.
12. Анцыферов С.С. Проблемы искусственного интеллекта // Проблемы искусственного интеллекта. — 2015. — № 0(1). — С.5-12.

ФОРМИРОВАНИЕ ПЛАНА ЗАКУПОК АВТОМОБИЛЬНЫХ ЗАПЧАСТЕЙ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ И ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ

Д.Д.Барышев, О.И.Пятковский

Алтайский государственный технический университет им. И. И. Ползунова
пр. Ленина 46, кафедра ИСЭ, т. 8(3852) 29-08-70, denis.baryshev@bk.ru

В условиях современных экономических отношений существенно возрастает сложность управления предприятием. Это объясняется увеличением его самостоятельности и развитием преимущественно горизонтальных связей с внешним окружением - поставщиками, потребителями, налоговыми инспекциями, банками. Для выработки рациональных решений требуется всесторонняя информация о производственно-хозяйственной, финансовой, снабженческо-сбытовой деятельности и внешних связях хозяйствующего субъекта.

Актуальность научных исследований и разработки автоматизированной аналитической системы формирования плана закупок продукции обусловлена постоянно увеличивающимся спросом с одной стороны и усиливающейся конкуренцией между различными производителями с другой. Но в тоже время нельзя допускать возникновения дефицита на складах. Выигрывает то предприятие, которое сможет наладить эффективные поставки. Для этого ему необходимо правильно прогнозировать величину спроса в предстоящем периоде и сформировать оптимальный заказ на закупку товаров.

Для обеспечения поддержки принятия решений в таких условиях необходимо дальнейшее развитие комплексной автоматизации управленческой деятельности предприятия с использованием современных информационных технологий, математического моделирования и методов искусственного интеллекта.

Подходы к построению информационных систем анализа и диагностики состояния предприятия и прогнозирования спроса приведены в ряде работ. На российском рынке предлагается немало программных продуктов, относящихся к классу аналитических систем. Они различаются как по спектру задействованных показателей, так и по реализованным в них подходам к решению основных задач анализа деятельности предприятий. Однако, как показывает анализ, в данных разработках недостаточно используются современные математические методы, нейросетевые и гибридные модели, позволяющие с гораздо большей точностью и качеством осуществлять функции диагностики, анализа и прогнозирования экономических показателей.

Существует также большое количество зарубежных информационных систем диагностики деятельности предприятия и прогноза его показателей. Изучение данных разработок показывает, что основной их недостаток состоит в том, что они имеют самостоятельное значение и не связаны с действующими информационными системами предприятия. Это не позволяет учитывать динамику изменения информационной базы, отображать реальные процессы управления на предприятии.

При исследовании функций управления внутри предприятия выявляются также многочисленные неформализованные задачи по оценке, диагностике и прогнозированию состояния предприятия, группировки данных, извлечения знаний на разных уровнях управления. В связи с возрастанием сложности производственных систем, необходимостью качественного оперативного принятия решений, своевременного выполнения заказов потребителей число таких задач существенно возросло.

Определены требования к системе формирования плана закупок, в соответствии с которыми её основными компонентами следует считать:

- информационную систему, которая решает задачу сбора и первичной обработки информации;
- систему принятия решений, которая, на основании сведений об объекте, установленных моделях, критериях качества и приоритетах, производит вывод правил принятия решений;
- систему ввода/вывода запросов, которая обеспечивает представление данных, полученных из системы, в виде, необходимом ЛПР.

На основе системного подхода представлен алгоритм формирования плана закупок с применением нейросетевых технологий и генетического алгоритма.

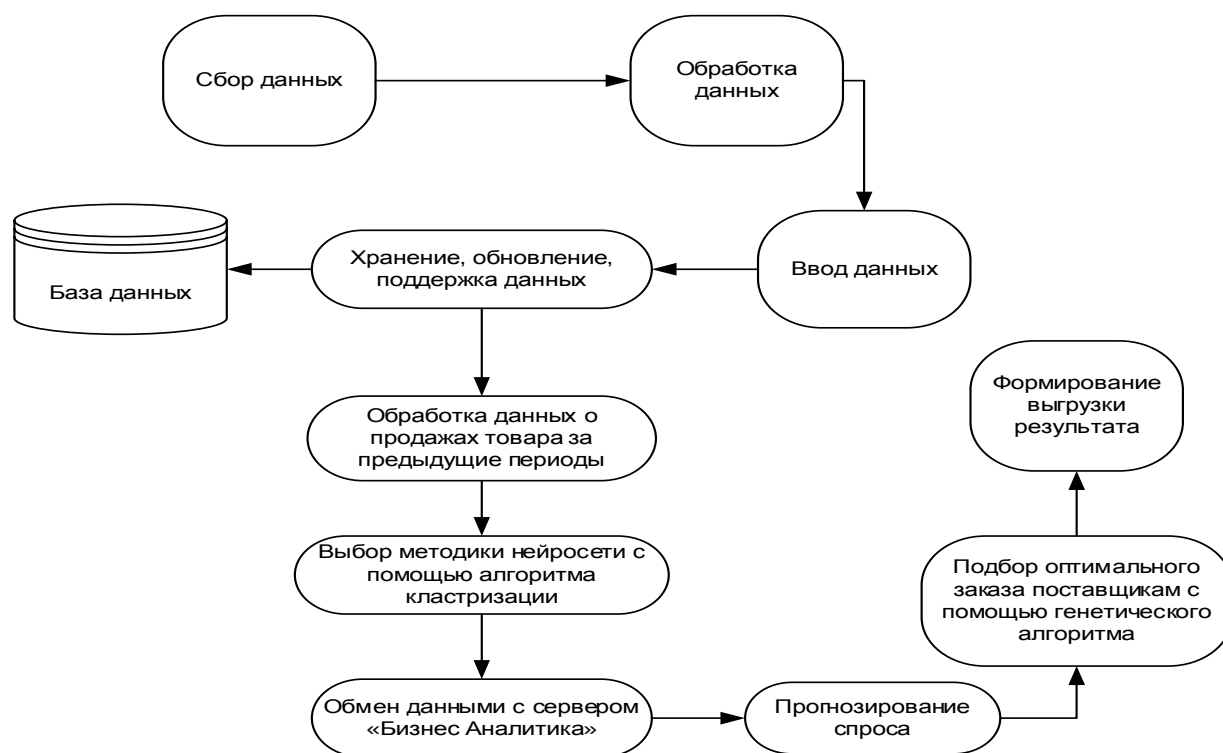


Рис.1. Алгоритм формирования плана закупок.

После сбора и обработки данных о заказах покупателей данные в формате CSV загружаются в локальную БД АИС, далее система обрабатывает данные о продажах за предыдущие периоды, алгоритмом кластеризации определяет методику прогнозирования для каждого товара, отправляет данные в формате XML на сервер «Бизнес аналитика», откуда получает значение прогнозируемого спроса, на основании этих данных формируется оптимальный заказ, который подбирается генетическим алгоритмом.

У каждого товара спрос изменяется по своему закону, для того чтобы составить прогноз спроса на товары, необходимо разбить товары на группы – кластеры, каждому кластеру будет соответствовать своя методика определения спроса. Для распределения товаров на группы использовался алгоритм k-средних. С помощью данного алгоритма были получены 5 обучающих выборок – центроидов кластеров, именно эти выборки и использовались для обучения нейронных сетей.

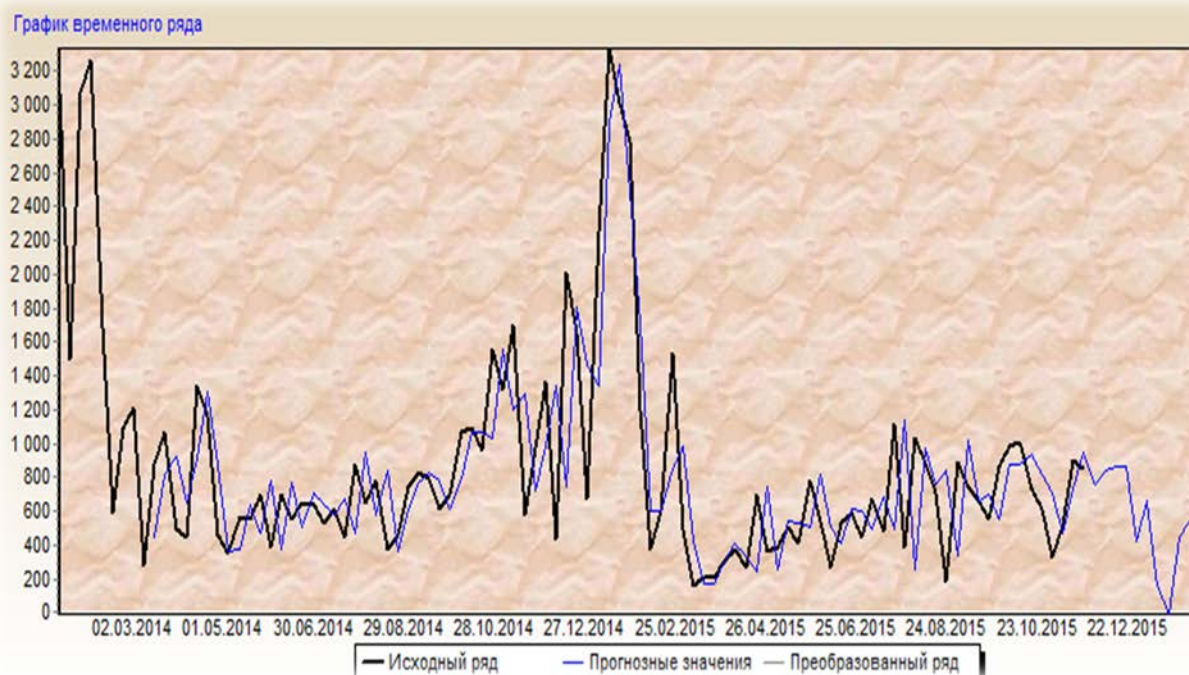


Рис.2. График значений для выборки.

Графики значений обучающих выборок показывают, что спрос на товары изменяется скачками, также можно заметить, что зимой спрос на товары значительно выше.

Для обучения методик «Бизнес-Аналитика» был проведен ряд экспериментов по подбору параметров нейронных сетей.

В результате множества экспериментов были получены параметры для обучения каждой выборки.

Таблица

Обучение нейронных сетей

Номер выборки	Метод обучения	Глубина погружения	Структура нейронной сети	Метод фильтрации	Параметры метода	Ошибка обобщения
1	kParTan	8	4-2-1	Фильтр Хемминга		0,141
2	kParTan	12	5-2-1	Вейвлет-преобразование	Глубина=2, порядок=4	0,087
3	kParTan	10	4-2-1	Медианное сглаживание	Ширина = 3	0,147
4	kParTan	10	5-3-1	Медианное сглаживание	Ширина = 5	0,067
5	kParTan	10	5-2-1	Вейвлет-преобразование	Глубина=3, порядок=6	0,098

Особенностью нейросетевых моделей являются их внутренние регуляризирующие свойства, позволяющие получать малые ошибки обобщения. Полезность таких свойств нейронных сетей проявляется в ситуациях, когда экспериментальные данные о системе содержат внутреннюю избыточность. Избыточность позволяет представить совокупность данных моделью, содержащей меньшее число параметров, чем имеется данных. Таким образом, нейросетевая модель сжимает экспериментальную информацию, устраняя шумовые компоненты и подчеркивая непрерывные, гладкие зависимости.

Для решения задачи прогнозирования спроса в зависимости от различных факторов в аналитической системе "Бизнес-Аналитик" были построены соответствующие модели на основе нейронных сетей. Для автоматизированной загрузки исходных данных и выгрузки рассчитанного по методике результата были применены web-сервисы "Бизнес Аналитика". В качестве основного формата обмена данными используется формат XML. На сервер отправляются данные о продажах, а в ответном сообщении приходят данные о спросе на товар.

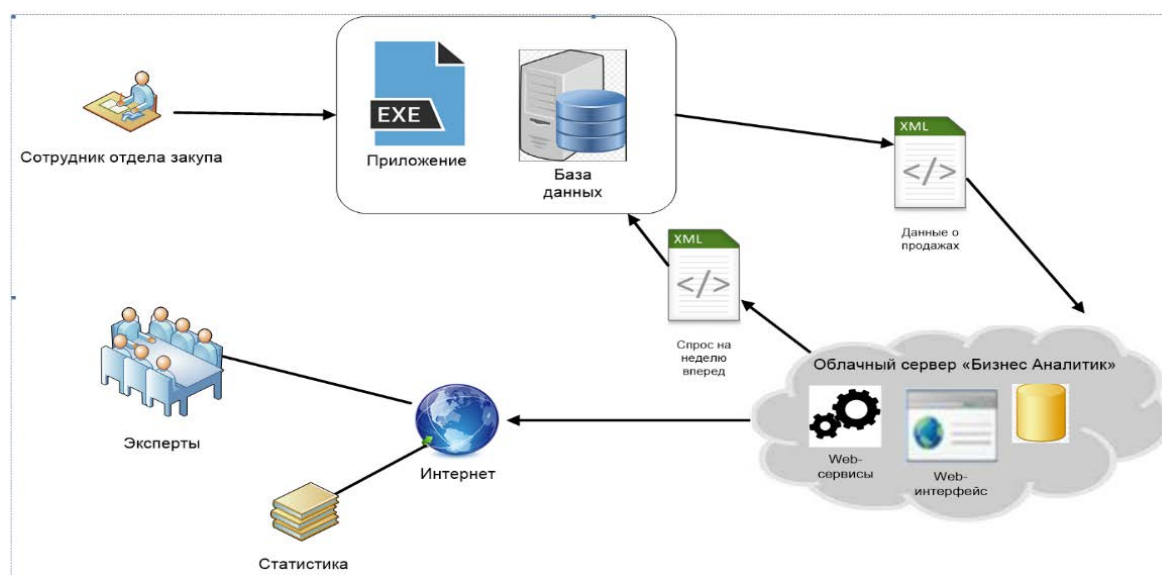


Рис.3. Интеграция с АИС «Бизнес-Аналитик».

В качестве основного формата обмена данными используется формат XML. Результат, выданный системой, отображается на форме.

Рис.4. Настройка параметров генетического алгоритма.

После работы с системой «Бизнес-Аналитик» АИС запускает генетический алгоритм, который начинает работать с вариантами заказов поставщикам, после этого прибыль самого выгодного заказа при каждом поколении генетического алгоритма выводится на экран. В завершении работы программы выводится оптимальный план закупа.

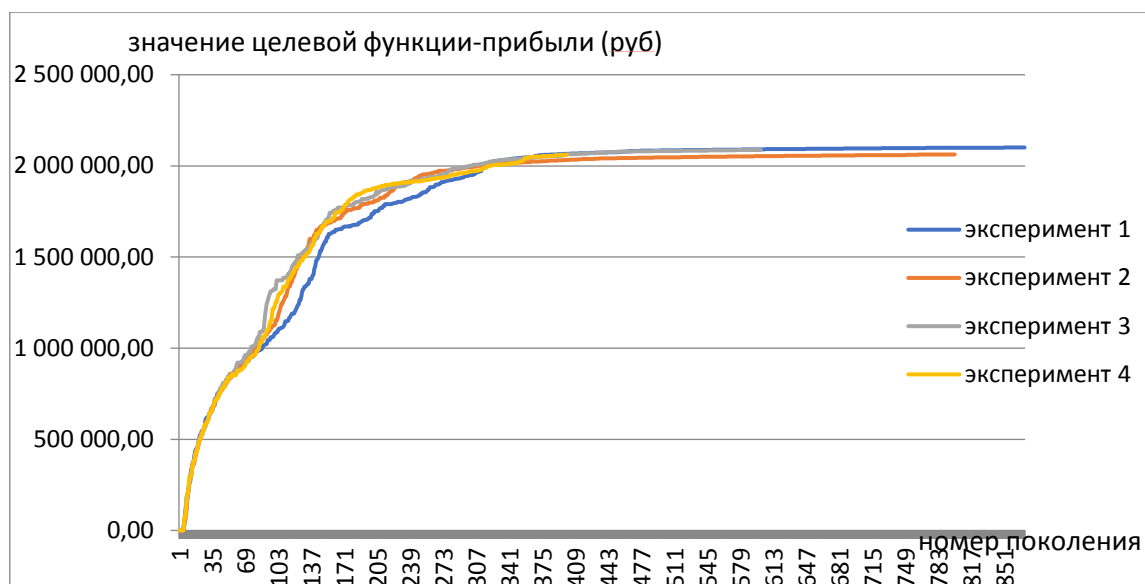


Рис.4. График изменения целевой функции генетического алгоритма в зависимости от изменения числа поколений.

Анализ экспериментов подтверждает эффективность разработанной модели, алгоритмов и технологий. Снижение ошибки прогноза происходит за счет ввода корректировки прогноза на основе настроенной базы знаний, отображающей знания и опыт экспертов.

Список литературы

1. Гаврилова, Т.А. Базы знаний интеллектуальных систем / Т. А. Гаврилова, В.Ф. Хорошевский – СПб.: Питер, 2000. – 384с.
2. Горбань А.Н. Обучение нейронных сетей / А.Н. Горбань. - М.: изд. СССР-США СП "ParaGraph", 1990. - 159с.
3. Миркес Е.М. Нейрокомпьютер. Проект стандарта / Е.М. Миркес. - Новосибирск: Наука. Сибирское предприятие РАН, 1999. - 337с.
4. Пятковский О.И. Интеллектуальные компоненты автоматизированных информационных систем управления предприятием: Монография / О.И. Пятковский. - Барнаул: изд-во АлтГТУ, 1999. - 351с.

СИММЕТРИЯ В СТАТИСТИЧЕСКИХ СТРУКТУРАХ ТРАНСКРИПТОМА И ГЕНОМА ХВОЙНЫХ ВИДОВ²

В.В.Бирюков¹, С.В.Новикова¹, Ю.А. Путинцева¹, М.Г.Садовский^{1,2}, К.В. Крутовский^{1,3,4,5}

¹Сибирский федеральный университет,
Институт фундаментальной биологии и биотехнологии, базовая каф. защиты и современных технологий мониторинга лесов

²Институт вычислительного моделирования СО РАН, msad@icm.krasn.ru

³Геттингенский университет им. Георга-Августа

⁴Институт общей генетики им. Н. И. Вавилова РАН

⁵Техасский A&M университет

Результаты *de novo* секвенирования транскриптов лиственницы сибирской (*Larix sibirica* Ledeb.) и сосны сибирской кедровой (*Pinus sibirica* Du Tour) были впервые получены в Лаборатории лесной геномики СФУ в рамках проекта полногеномного секвенирования хвойных бореальных лесов под руководством проф. К.В. Крутовского [1]. Данные секвенирования лиственницы были получены на приборах Illumina с использованием парно-концевых библиотек размером 250 – 400 нуклеотидных оснований (н.о.). Для сборки был использован транскриптомный *de novo* ассемблер Trinity, основанный на графах де Брёйна [2]. Характеристики полученных сборок представлены в табл. 1.

Таблица 1

Основные характеристики полученных сборок

Показатель	<i>L. sibirica</i>	<i>P. sibirica</i>
Число контигов (≥ 0 н.о.)	43 686	1 083 468
Число контигов (≥ 1000 н.о.)	6 919	79 456
Длина максимального контига, н.о.	8 512	15 689
Общая длина (≥ 0 н.о.)	26 493 048	501 364 133
Общая длина (≥ 1000 н.о.)	11 996 463	148 545 280
N50, н.о.	878	503
N90, н.о.	259	234
GC, %	43,36	39,67

В связи с большим количеством собранных контигов транскриптома камбия сосны сибирской была проведена структурная аннотация при помощи программного обеспечения Transdecoder, которая позволила сократить количество исследуемых данных. TransDecoder – инструмент для предсказания открытых рамок считывания (ORF) в результатах сборки [3]. В результате запуск программы со стандартными значениями параметров позволил оставить только те последовательности, которые имеют целостную структуру мРНК. Таким образом, для дальнейших исследований нами были отобраны 123287 контигов транскриптома камбия сосны сибирской.

Было проведено сравнение нуклеотидных последовательностей полученных сборок с базой данных секвенированных нуклеиновых кислот с помощью программы локального выравнивания BLAST. Большинство собранных транскриптов имеют

² Работа выполнена при финансовой поддержке гранта № 14.Y26.31.0004 Правительства РФ «Геномные исследования основных бореальных лесобразующих хвойных видов и их наиболее опасных патогенов в Российской Федерации».

гомологи в нуклеотидной базе NCBI, представляющие собой фрагменты мРНК представителей хвойных.

На основе распределения длин контигов, для первичного статистического анализа полученных транскриптомов отбирались последовательности, длина которых составляла не менее 2000 н.о.: 1436 и 5628 последовательностей для транскриптомов лиственницы и сосны, соответственно). Затем для каждой последовательности составлялся частотный словарь триплетов. Под частотным словарем каждого фрагмента следует понимать список из 64 триплетов с указанием частоты каждого. Таким образом, каждый элемент транскриптома преобразовывался в точку в 64-мерном (Евклидовом) пространстве, где координатами являются триплеты, а их значениями — частоты [4].

В данной работе были составлены и изучены частотные словари четырех типов. Первый тип частотного словаря составлен в результате сдвига рамки считывания (3 н.о.) на 1 нуклеотид, начиная от 1-го нуклеотида последовательности. Второй, третий и четвертый частотные словари получены в результате сдвига рамки считывания на 3 нуклеотида от 1-ой, 2-ой и 3-ей позиции, соответственно. В первом случае рамки считывания пересекались, в других — нет. На рис. 1 схематично представлены частотные словари, полученные в работе.

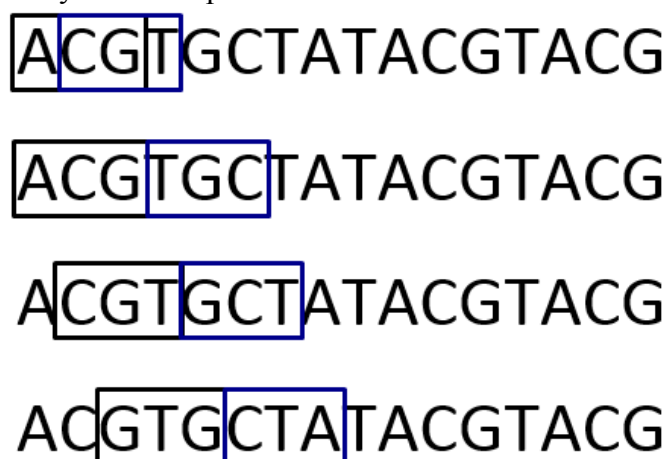


Рис. 1. Различные типы частотных словарей

Изучалось разбиение множества указанных элементов транскриптома на классы методом динамических ядер при помощи программы ViDaExpert: строилось разбиение на 2, 3, 4, и 5 классов [5]. Для исключения влияния линейной связи, создаваемой тем, что сумма всех частот всех триплетов равна 1, один из триплетов с наименьшим стандартным отклонением исключался из рассмотрения. Такой выбор обусловлен тем, что такой триплет давал наименьший вклад в различимости частотных словарей.

Стоит отметить, что построение классификации методом динамических ядер всегда начинается со случайного разбиения точек исходного множества на два класса и сходимость в серии различных испытаний к одному и тому же распределению вовсе не гарантирована. Для проверки достоверности получаемой кластеризации (устойчивости распределения точек пространства в один и тот же кластер независимо от выборки и стартовых точек) было проведено достаточно большое количество делений имеющихся последовательностей методом динамических ядер. В случае транскриптома лиственницы устойчивость получаемого разбиения была очень высокой — в серии из пяти сотен реализаций МДЯ все пятьсот раз в один и тот же класс попадало более 95% всех последовательностей этого класса. Более того, те последовательности, которые меняли свою принадлежность к классу, во-первых, делали это тоже сравнительно редко — из трёхсот реализаций лишь в 35 они переходили в противоположный класс, а

во-вторых, группа переходящих последовательностей была всегда одной и той же. Это доказывает, что построенная классификация была очень устойчивой. Кроме того, высокой была и различимость классов: расстояние между центрами классов (0,0352) было больше каждого из радиусов полученных классов (0,0269 и 0,0272). Транскриптом сосны сибирской также показал высокий уровень различимости при делении на два и три кластера (порядка 98 % последовательностей сохраняли свою принадлежность к тому или иному классу в обоих случаях).

На следующих этапах работы рассматривалась только кластеризация на 2 класса, т.к. она является наиболее устойчивой и имеет высокую различимость классов. Для визуализации результатов кластеризации последовательностей транскриптома лиственницы был использован метод упругих карт (вид во внутренних координатах) – см. рис. 2. Было предположено, что полученные два кластера соответствуют «минус-цепи» и «плюс-цепи» ДНК. Для проверки данной гипотезы было использовано два подхода: расчет величины невязки и проверка с помощью GenBank.

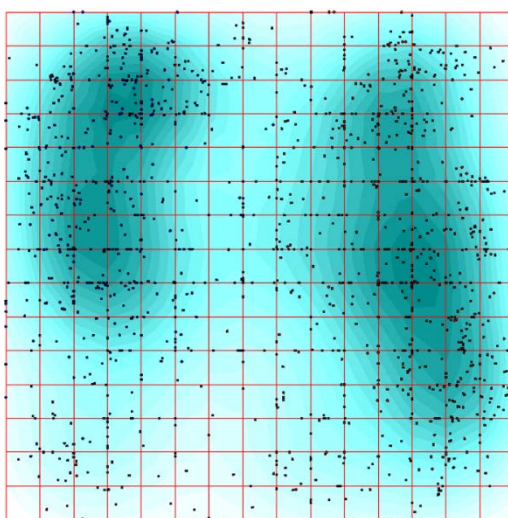


Рис. 2. Результаты кластеризации транскриптома лиственницы на 2 класса

Хорошо известно, что реальные нуклеотидные последовательности обладают важными комбинаторными свойствами: для них выполняются (с той или иной точностью) т.н. первое и второе правила Чаргаффа [6 – 8]. Первое правило Чаргаффа гласит, что в молекулах ДНК количество тимина (Т) равно количеству аденина (А), а количество гуанина (G) равно количеству цитозина (С). Второе правило Чаргаффа гласит, что в пределах одной цепочки ДНК доля нуклеотидов А близка к доле нуклеотидов Т (и, соответственно, близки доли нуклеотидов G и С).

Обобщённое второе правило Чаргаффа гласит, что в реальных последовательностях близкими являются частоты двух олигонуклеотидов, читающиеся одинаково в противоположных направлениях, с учётом замены нуклеотидов по правилу комплементарности, т.н. комплементарный палиндром. Хорошо известно, что различные таксономические группы организмов различаются по тому, насколько точно выполняется второе правило Чаргаффа для их геномов (GC-состав). Одна из возможных мер μ_3 точности выполнения определяется так (нижний индекс подчёркивает тот факт, что мера μ_3 определена для триплетов):

$$\mu_3 = \|\Omega\|^{-1} \sqrt{\sum_{\omega \in \Omega} (f_{\omega} - f_{\bar{\omega}})^2} \quad (1)$$

где ω и ϖ — пара триплетов, составляющих комплементарный палиндром, Ω — множество таких пар, символ $||\cdot||$ обозначает мощность множества, а f_ω — частоту слова ω . Характерная величина показателя (1) для генетических систем разной природы варьирует от 0.05 до 0.002 [6].

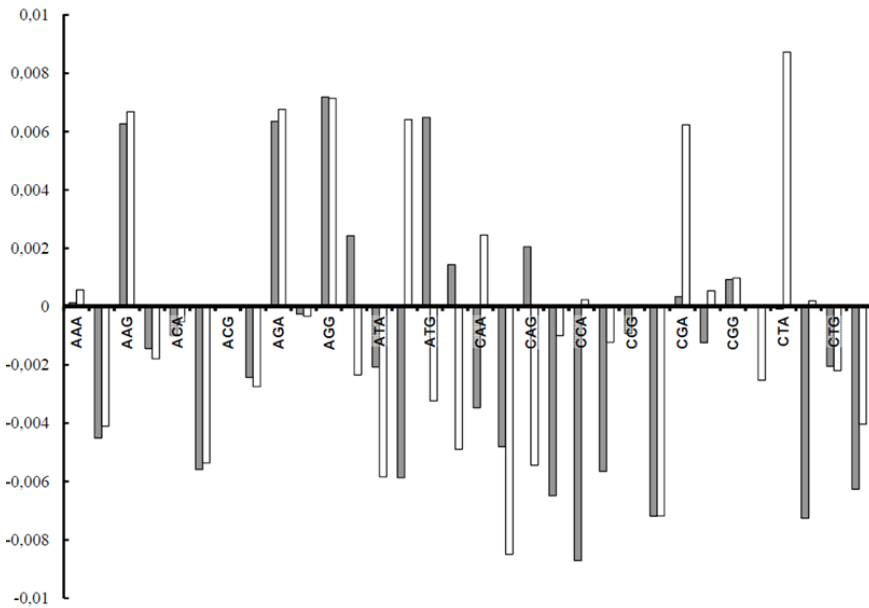


Рис. 3. Величина невязки внутри классов (серым цветом обозначен 1 класс, белым – 2 класс)

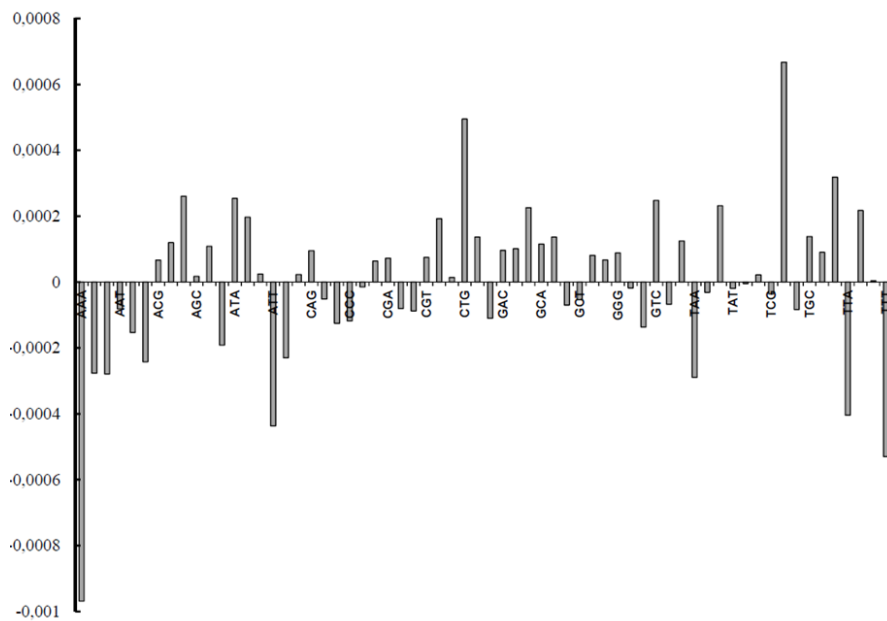


Рис. 4. Величина невязки между классами

Для ядра каждого из двух полученных методом динамических ядер классов была определена величина невязки (1); при этом она определялась по 32 парам триплетов, составляющих комплементарные палиндромы. Результаты определения показаны на рис. 3. Затем определялась невязка (1) между центрами двух классов — см. рис. 4. Степень различия невязки, определяемой между двумя классами, и внутри каждого из классов, оказалась достаточно высокой (более порядка). Так невязка в 1 и 2 классе

составила $7,788 \times 10^{-4}$ и $7,761 \times 10^{-4}$, соответственно, в то время как невязка между классами составила $2,919 \times 10^{-5}$.

Наблюдаемый эффект может быть обусловлен особенностью построения транскриптома: контиги в этом наборе разнонаправлены, т.е. представляют собой смесь $5' \rightarrow 3'$ и $3' \rightarrow 5'$ фрагментов. Хотя РНК выделялась с помощью поли-А конца, т.е. в одном направлении, при синтезе кДНК и дальнейшем её секвенировании и сборке в консенсусные контиги могут быть ориентированы в любом направлении.

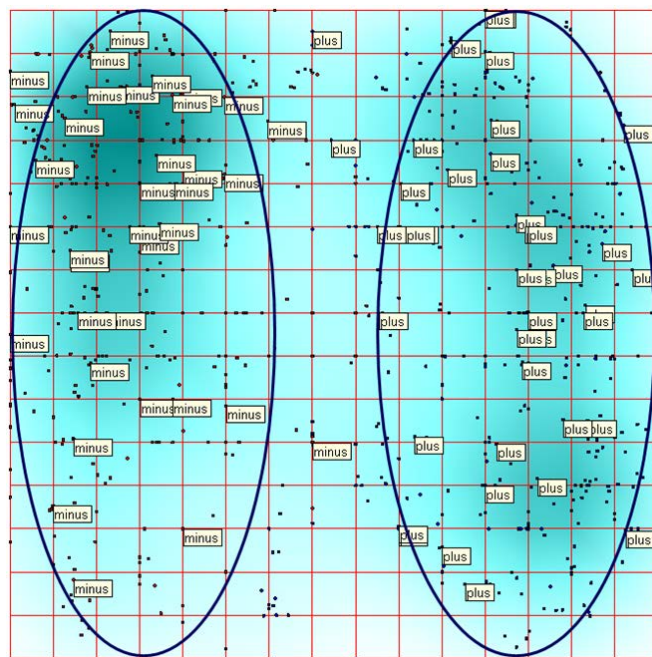


Рис. 5. Результаты кластеризации на 2 класса последовательностей транскриптома лиственницы с указанием информации из GenBank

Метод проверки при помощи BLAST заключался в следующем. Все контиги выборки были выровнены на нуклеотидную базу GenBank с E-value, равным 10^{-5} . В результате, для 1436 последовательностей было обнаружено 759 соответствий, из них «плюс-цепи» соответствовало 377 последовательностей, «минус-цепи» — 381. Важно отметить, что для каждой последовательности было получено 5 лучших выравниваний, из которых мы брали направление по самому первому, в большинстве случаев направление у всех 5 совпадало, но были и исключения. Для 677 последовательностей совпадений в GenBank не нашлось. Для последовательностей, найденных в базе, было проведено повторное разделение последовательностей на два класса. Результат кластеризации представлен на упругой карте, представленной на рис. 5. Для удобства восприятия данной иллюстрации направления считывания указаны только для 20 % последовательностей.

Аналогичная работа была проделана для генома лиственницы. После кластеризации геномных данных на два класса были посчитаны невязки внутри классов: $2,55 \times 10^{-5}$ и $1,71 \times 10^{-5}$ между классами: $1,46 \times 10^{-3}$. При этом для генома наблюдалась картина невязок, противоположная той, которая наблюдалась для транскриптома. Это позволяет предположить, что для этой выборки в каждый класс попали контиги, принадлежащие обоим стрендам — результат, противоположный распределению контигов при кластеризации транскриптомных данных.

Кластеризация частотных словарей второго, третьего и четвертого типов транскриптомных данных лиственницы и сосны на классы дала похожие результаты. На

упругих картах (во внутренних координатах), представленных на рис. 6, хорошо видны 6 кластеров, соответствующих трем кластерам с контигами с «плюс-цепи» и трем кластерам с фрагментами с «минус-цепи» ДНК. Используя частотные словари 3-го и 4-го типов, был проверен факт перехода контигов между кластерами. Контиги, находящиеся в одном кластере, представляют другой кластер контигов упругой карты, построенной на основе частотных словарей другого типа (рис. 7). Переход осуществляется полностью без разделения. Тем самым было показано, что образованные 6 кластеров – вершины фазовых треугольников, образуемые в результате сдвига рамки считывания [9]. Косвенным подтверждением этого служит то, что при кластеризации частотных словарей контигов, принадлежащих только одному стренду ДНК, образуется трехкластерная структура, соответствующая одному из фазовых треугольников.

В статье [9] также было показано, что геномные данные при кластеризации образуют 7-кластерную структуру, где седьмому центральному кластеру соответствуют некодирующие участки генома. Транскриптом не содержит таковых, и, следовательно, получаемая структура имеет только 6 кластеров. Используя метод главных компонент, были получены трехмерные изображения данных структур, каждый кластер которой представляет собой вершину октаэдра (см. рис. 8). Такое расположение объясняется GC-составом, как было показано в [9].

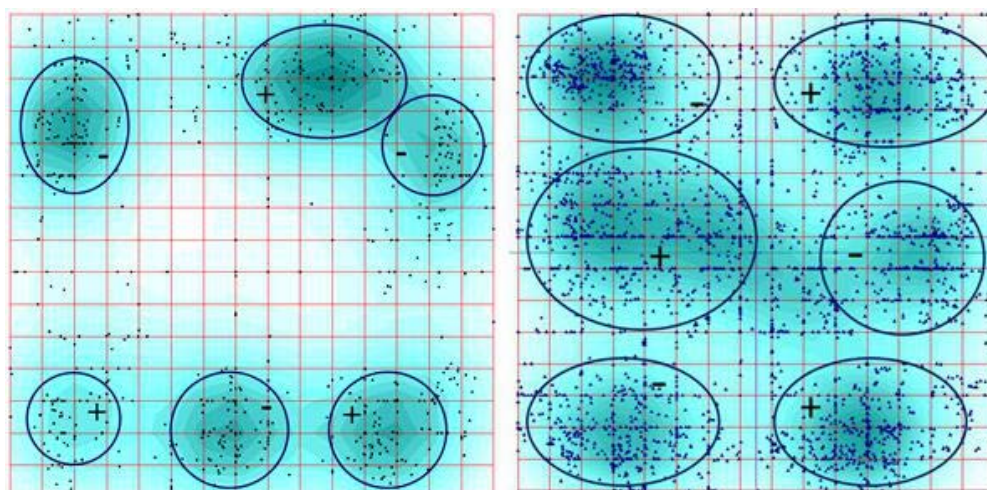


Рис. 6. Результаты кластеризации частотных словарей транскриптома 2 типа для лиственницы сибирской (слева) и для сосны сибирской кедровой (справа)

При дальнейшем анализе транскриптомных данных словари частот встречаемости триплетов строились уже не для контигов, а для фрагментов внутри них. Последовательность разбивалась на фрагменты по 500 н.о. со сдвигом стартовой позиции следующего фрагмента на десять нуклеотидов. Наблюдалась устойчивое распределение фрагментов на классы таким образом, что принадлежность к тому или иному классу соответствовала фазе — смещению стартового нуклеотида фрагмента относительно начала кодирующего участка транскрипта. Иными словами, в первый кластер попали фрагменты, у которых фаза совпадает с первым нуклеотидом старт-кодона последовательности, во второй кластер — совпадение фазы со вторым нуклеотидом, а в третий — совпадение с третьим нуклеотидом иницирующего кодона. Такие кластеры при трехмерной визуализации методом главных компонент отображаются как вершины правильного треугольника (рис. 9). Появление четвертого кластера, находящегося посередине, обусловлено наличием фрагментов, не вошедших в кодирующий регион транскрипта — нетранслируемые области на концах мРНК (3'-UTR и 5'-UTR).

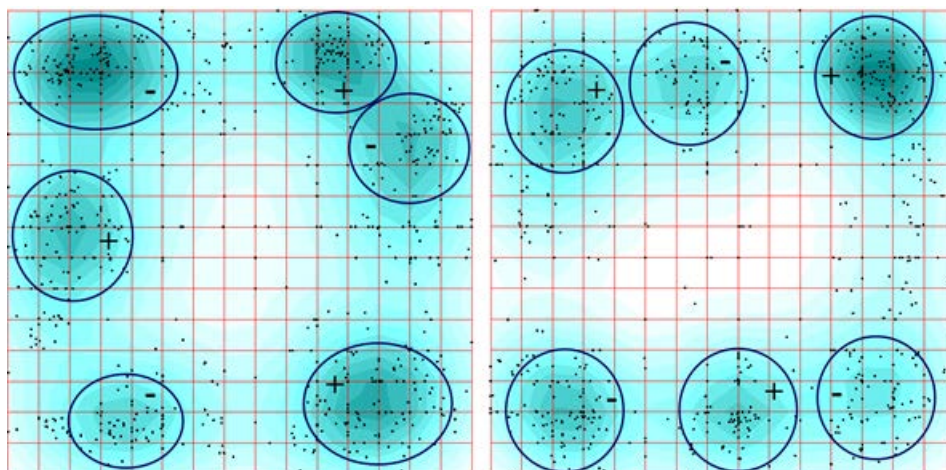


Рис. 7. Результаты кластеризации частотных словарей транскриптома лиственницы сибирской 3 и 4 типов

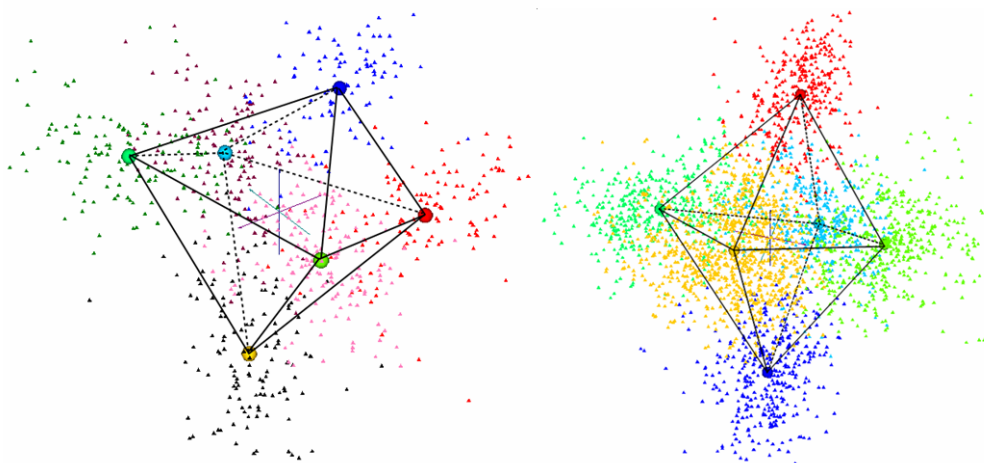


Рис. 8. 6-кластерные октаэдральные структуры транскриптомов лиственницы (слева) и сосны (справа)

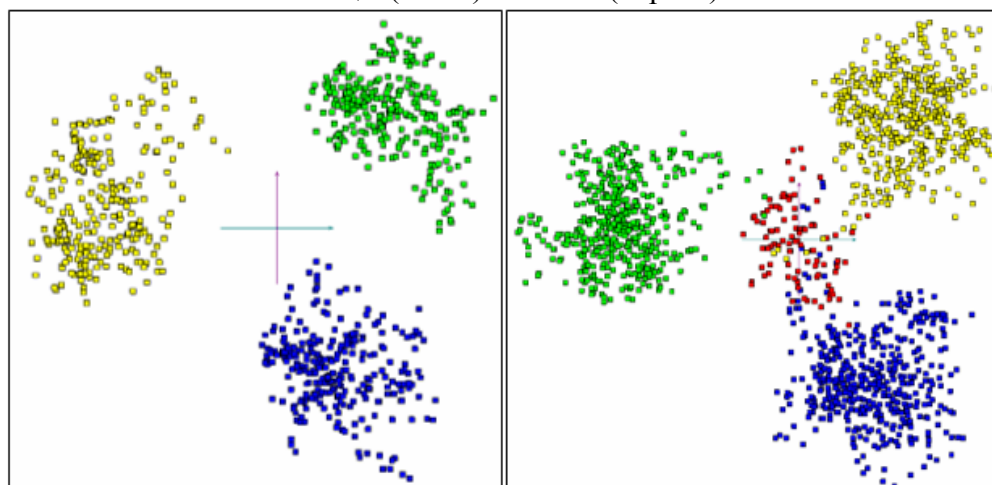


Рис. 9. Визуализация фрагментарных частотных словарей для лиственницы сибирской (слева) и сосны сибирской кедровой (справа). Разными цветами обозначены фазы фрагментов.

На транскриптомных данных двух видов хвойных было показано, что кластеризация частотных словарей триплетов методом динамических ядер позволяет разделить транскриптомные последовательности на структурно различные группы. Установлено, что такая кластеризация в качестве первого шага обработки транскриптомных данных может существенно упростить дальнейшую работу, например, при определении направления считывания транскриптов. При кластеризации частотных словарей первого типа на две группы, последовательности разбиваются согласно своей стренд-специфичности. Однако при кластеризации генома такого разбиения не наблюдается — в оба кластера попадают последовательности из разных стрендов.

Построение частотных словарей 2, 3 и 4-го типов позволяет разделить набор транскриптомных данных в соответствии с их относительной фазой. Такой результат указывает на то, что транслируемая область мРНК имеет совершенно неслучайную структуру, и описанный метод может быть применен для поиска границ транслируемой области, либо для перепроверки качества ассемблированных транскриптов. Полученная в результате кластеризации шестикластерная структура транскриптомов весьма схожа с полученной ранее семикластерной структурой генома бактерий, за исключением отсутствия у транскриптома центрального кластера с некодирующими последовательностями. При этом расположение данных кластеров соответствует вершинам октаэдра, положение которых объясняется GC-составом последовательности [9].

Список литературы

1. Крутовский К.В., Орешкова Н.В., Путинцева Ю.А., и др. // Сибирский лесной журнал. — 2014. — Т. 1. — № 4. — С. 79 – 83.
2. Grabherr M., Haas B., Yassour M., et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome // *Nature Biotechnology*. — 2011. — V. 29. — P. 644 – 652.
3. Haas B., Papanicolaou A., Yassour M., et al. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity // *Nat Protoc*. — 2013. — 8(8). — 10.
4. Gorban A.N., Popova T.G., Sadovsky M.G., Wunsch D.C. Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts // *Smart Engineering System Design*. — 2001. — 11. — P. 657 – 663
5. Gorban A.N., Pitenko A., Zinovyev A. ViDaExpert: user-friendly tool for nonlinear visualization and analysis of multidimensional vectorial data. E-print: <https://arxiv.org/abs/1406.5550>
6. Гребнев Я.В., Садовский М.Г. Второе правило Чаргаффа и симметрия геномов // *Фундаментальные исследования*. — 2014. — Т. 12. — №. 5. — С. 965 – 968.
7. Mascher M., Schubert I., Scholz U., Friedel S. Patterns of nucleotide asymmetries in plant and animal genomes // *BioSystems*. — 2013. — 111(3). — P. 181 – 189
8. Afreixo V., Rodrigues J.M., Bastos C.A. Analysis of single-strand exceptional word symmetry in the human genome: new measures // *Biostatistics*. — 2015. — 16(2). — Pp. 209 – 221
9. Gorban A.N., Zinovyev A.Y., Popova T.G. Seven clusters in genomic triplet distributions // *In Silico Biol*. — 2003. — 3(4). — P. 471 – 482

ВЛИЯНИЕ ТЕХНИК САМОРЕГУЛЯЦИИ НА ПОКАЗАТЕЛИ ЗДОРОВЬЯ СТУДЕНТОВ

Н.М. Богатов, А.Н. Богатова, В.С. Лахин, С.А. Сухих

Кубанский государственный университет, *bogatov@phys.kubsu.ru*

Организм человека является открытой энергетической системой, которая вступает во взаимодействие с внешней средой. Агрессивное воздействие внешней среды проявляется в возрастании психофизиологических, информационных нагрузок при адаптации к современным социально-экономическим условиям и техногенным факторам, разрушающим экологию и биосферу. Следствием этого является повышение биоэнергетических затрат организма, приводящее к истощению его ресурсов, появлению хронической усталости, хронического стресса, различных психологических и соматических расстройств. Саморегуляция является важнейшим фактором достижения физиологической, психологической и социальной адаптации. Саморегуляция реализуется как на уровне отдельных функциональных систем организма, так и посредством их взаимодействия [1], в котором важнейшую роль играют нейрогуморальные реакции. Способность индивидуума к саморегуляции посредством управления психоэнергетическим потенциалом является индивидуальной характеристикой человека. Положительное влияние психотехник на саморегуляцию функциональных состояний человека продемонстрировано в работах [2 – 5].

Студенты – это особая социальная группа, определяющая в перспективе трудовой и оборонительный потенциал страны, воспроизводство и здоровье будущих поколений. Условия труда и жизни студентов характеризуются повышенными психоэмоциональной, умственной нагрузками, а также необходимостью адаптации к изменяющимся условиям обучения и проживания. Сохранение здоровья студентов является составной частью общегосударственной проблемы сохранения и укрепления здоровья населения РФ.

Цель исследования – изучение влияния техник саморегуляции на функциональное состояние здоровья студентов.

В процессе эксперимента достигалось воздействие сознания субъекта на энергоинформационный обмен в организме. Информационное воздействие на орган или энергетический центр происходило во внутреннем пространстве субъекта. В эксперименте участвовали респонденты в возрасте 21 – 23 года в количестве 50 человек. Участники эксперимента не обладали специальной подготовкой по управлению сознанием.

Методика исследования включала три этапа: 1 – определение начального функционального и психоэмоционального состояния испытуемых, 2 – методически организованный сеанс саморегуляции, 3 – определение функционального и психоэмоционального состояния испытуемых после сеанса саморегуляции.

На первом и третьем этапах функциональное состояние организма испытуемых оценивалось методами функциональной диагностики, основанными на измерении электрических потенциалов с помощью программно-аппаратного комплекса SilverStep. Комплекс SilverStep, используя анализ вариабельности сердечного ритма, оценивает состояние регуляторных систем организма, в частности комплексный показатель здоровья, адаптацию, вегетативное равновесие нервной системы, спектрограмму ритмов головного мозга, нейрогуморальную регуляцию, психоэмоциональное состояние, энергетический баланс и др.

Сеансы саморегуляции проводились 1 раз в неделю в течение весеннего учебного семестра, длительность сеансов составляла 20 ÷ 30 минут. Использовались следующие психотехники: абдоминальное (брюшное) дыхание, медитация на 7 чакрах со звуковым

сопровождением (очищающая мантра). В исследовании приняли участие 22 студента (13 юношей, 9 девушек) в возрасте 20-22 года.

Сравним результаты измерений функциональных показателей здоровья студентов до и после сеансов саморегуляции. На рис.1 показана доля студентов, у которых наблюдалось увеличение функциональных и психоэмоциональных показателей, определяемых программой SilverStep PULSE. Из рис.1А следует, что в результате абдоминального дыхания улучшение всех показателей, представленных на диаграмме, наблюдалось более чем у 67% испытуемых, наилучший результат 80 % получен для показателя адаптации. Результаты сеанса медитации оказались ниже, чем абдоминального дыхания, (рис.1Б): доля испытуемых с улучшениями по всем пяти показателям составила более 60%, наилучший результат 73 % отмечен для показателя нейрогуморальной регуляции.

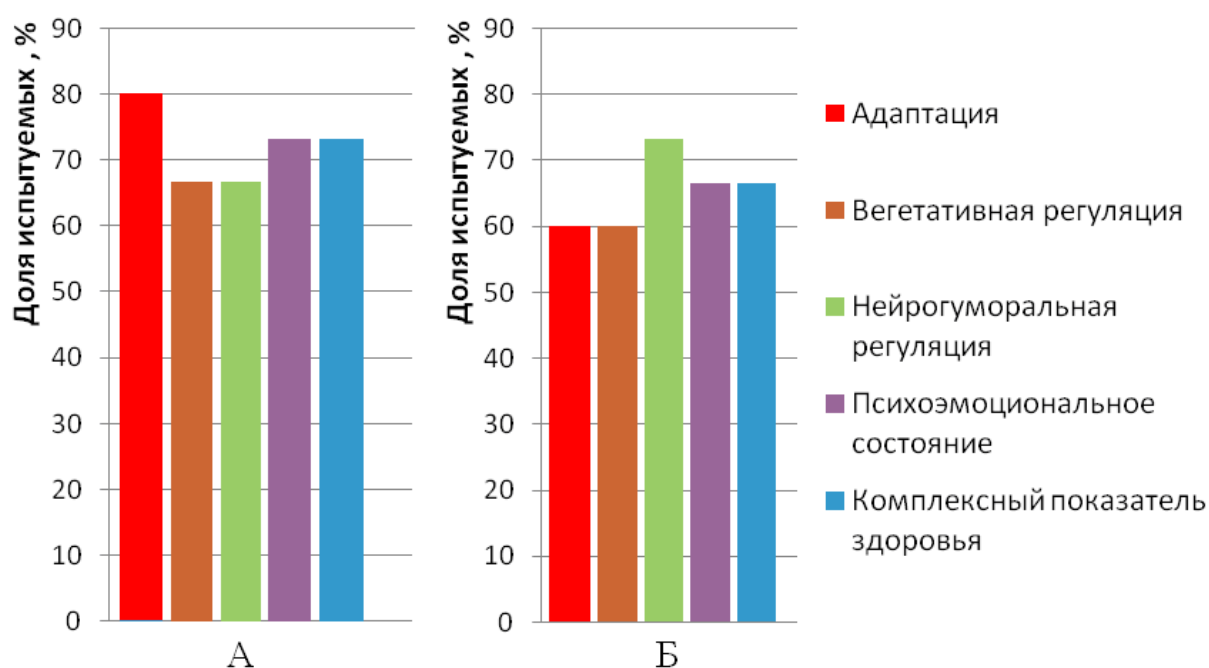


Рис.1. Доля испытуемых студентов с улучшениями показателей:
А – в результате абдоминального дыхания; Б – в результате медитации.

На рис.2 показана доля студентов, у которых наблюдалось увеличение показателей семи энергетических центров (чакр) и мощности ауры, определяемых программой SilverStep AURA. В таблице приведены названия чакр и их функциональная связь с организмом.

Таблица

Энергетические центры

№ чакры	Название	Функциональная связь
7	сахасрара	головной мозг
6	аджна	центральная нервная система
5	вишудха	эндокринная и лимфатическая системы
4	анаката	Сердечно-сосудистая и дыхательная системы
3	манипура	желудочно-кишечный тракт, печень, желчный пузырь
2	свадхистана	мочеполовая система, почки, обмен веществ
1	муладхара	жизненная сила, иммунитет

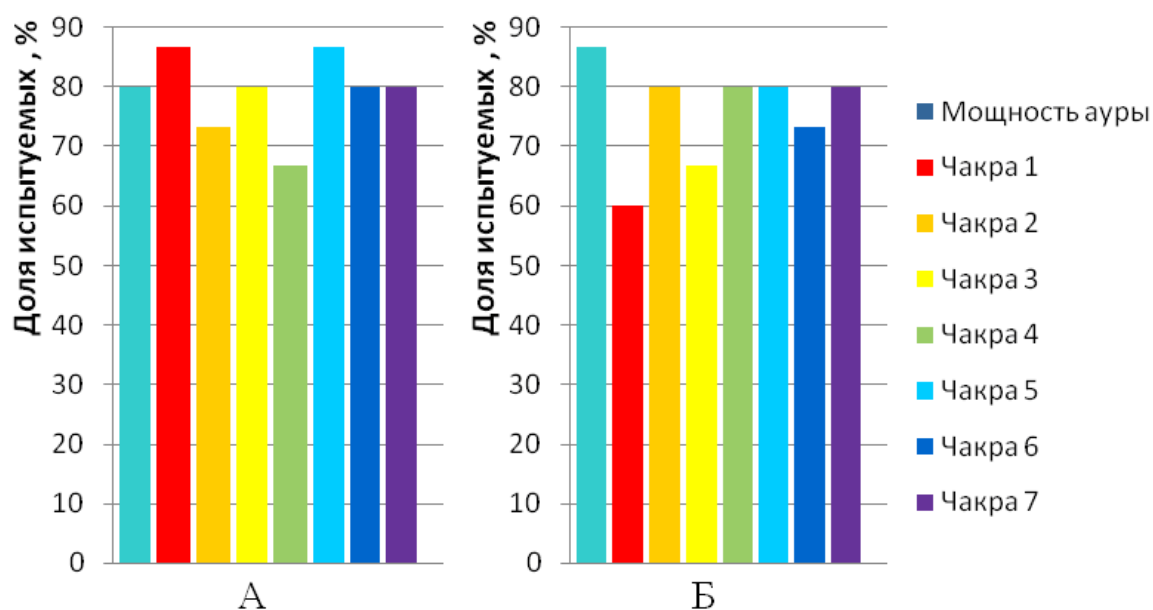


Рис.2. Доля испытуемых студентов с улучшениями показателей чакр:
А – в результате абдоминального дыхания; Б – в результате медитации.

Из рис.2А следует, что в результате абдоминального дыхания улучшение всех показателей, представленных на диаграмме, наблюдалось более чем у 67% участников исследования, наилучший результат, достигнутый у 87 % испытуемых, заключался в активизации чакр 1 и 5. Сеанс медитации дал иные результаты (рис.2Б): доля испытуемых с улучшениями по всем показателям составила более 60%, наилучший результат 87 % отмечен для показателя суммарной ауры.

Результаты исследования показывают, что практика правильного дыхания положительно влияет на функциональное и психоэмоциональное состояние здоровья студентов. Этот факт имеет физиологическое обоснование. В ответ на стресс напрягается группа мышц стенок брюшной полости. Если брюшные мышцы напряжены, то они давят на диафрагму при ее расширении во время вдоха, что уменьшает объем вдыхаемого воздуха и задерживает его в верхней части легких. При поверхностном дыхании может ощущаться нехватка кислорода, что вызывает напряжение других функциональных систем организма. Чтобы компенсировать недостаток кислорода учащается неглубокое дыхание, вместо того чтобы расслабить брюшные мышцы и дышать глубже.

Абдоминальное дыхание расслабляет мышцы, давящие на диафрагму, и дыхание замедляется. Несколько глубоких вдохов позволяют эффективно расслабиться и снизить напряжение функциональных систем организма. Наиболее значительными достижениями сеанса абдоминального дыхания являются улучшение показателей психоэмоционального состояния и комплексного показателя здоровья у 73 %, адаптации у 80 % студентов (рис.1А), энергетических центров, связанных с желудочно-кишечным трактом, печенью, желчным пузырем, центральной нервной системой, головным мозгом у 80 %, регенерацией, иммунитетом, эндокринной и лимфатической системами у 87 % студентов, участвовавших в эксперименте, (рис.2А).

Испытуемые после сеанса абдоминального дыхания ощущали более свободное дыхание и облегчение внутреннего состояния. В условиях постоянного стресса для сохранения функционального состояния здоровья абдоминальное дыхание следует регулярно использовать в повседневной жизни. Для самостоятельной практики дыхания не требуются особые условия, для усиления эффекта его можно выполнять одновременно с другими психотехниками.

Сеанс медитация на 7 чакрах сопровождается концентрацией внимания и последовательного его переключения. Концентрация внимания – это психический процесс и психическое состояние мысленной сосредоточенности на выбранных объектах при игнорировании других. С помощью концентрации внимания и визуализации некоторых образов можно управлять эндогенными процессами организма [6]. Медитация со звуковым сопровождением, в процессе которой достигается сосредоточенность на определенных энергетических центрах, позволяет направить работу нейрогуморальной системы в нужном направлении и, таким образом, добиться необходимого результата.

Результаты исследования показывают, что практика медитации оказывает положительное влияние на организм человека в целом и на его психофизиологическое состояние в частности. Наиболее значительными достижениями этого сеанса являются улучшение показателей психоэмоционального состояния и комплексного показателя здоровья у 67 %, нейрогуморальной регуляции у 73 % студентов (рис.1Б), энергетических центров, связанных с центральной нервной системой у 73 %, мочеполовой системой, почками, обменом веществ, сердечно-сосудистой и дыхательной системами, эндокринной и лимфатической системами, головным мозгом у 80 % испытуемых (рис.2Б).

Таким образом, полученные результаты свидетельствуют о снижении за счет использованной техники саморегуляции психоэмоционального и функционального напряжения организма, являющегося следствием усиления физиологических процессов, происходящих в организме в ответ на физические и интеллектуальные нагрузки. На этапе восстановления (сеансы саморегуляции), активность физиологических процессов снижается, что и обуславливает нормализацию функционального состояния организма.

Список литературы

1. Анохин П.К. Очерки по физиологии функциональных систем / П.К.Анохин. — М.: Медицина, 1975. — 448 с.
2. Богатов Н.М. Исследование процессов саморегуляции организма тепловизионным методом / Н.М. Богатов, А.Н. Богатова, С.А. Сухих // Нейроинформатика, её приложения и анализ данных: Материалы XXIV Всероссийского семинара, 30 сентября – 2 октября 2016 г. Красноярск. Красноярск: Институт вычислительного моделирования СО РАН, 2016. С. 9 – 12.
3. Sukhih S.A. Psychophysiological self-regulation in subject's bioenergetic safety / S.A. Sukhih, A.N. Bogatova, N.M. Bogatov // Приоритетные научные направления в XXI веке (том 5). Материалы Международной научно-практической конференции, 5 октября 2016 года г.Прага, Чехия. Прага: Vydavatel «Osviceni», Нефтекамск: РИО НИЦ «Мир науки», 2016. С. 46 – 53.
4. Bogatov N.M. Measuring bodily self-regulation precess with thermal imaging / N.M. Bogatov, A.N. Bogatova, S.A. Sukhih // Достижения современной науки (том 5). Материалы Международной научно-практической конференции, 7 октября 2016 года г.София, Болгария.София: Издателска Къща «СОРОС», Нефтекамск: РИО НИЦ «Мир науки», 2016. С. 13 – 18.
5. Богатов Н.М. Влияние психотехник на саморегуляцию функционального состояния человека / Н.М. Богатов, А.Н. Богатова, С.А. Сухих // Коллективная монография. Выпуск 5. Современные проблемы физики, биофизики и инфокоммуникационных технологий. – Краснодар: ЦНТИ, 2016. С. 5–16.
6. Вегетативные расстройства. Клиника, лечение, диагностика / Под ред. А. М. Вейна – М.: МИА, 2003. 752 с.

МЕТОД ГЛАВНЫХ И НЕЗАВИСИМЫХ КОМПОНЕНТ В АНАЛИЗЕ НАТУРНЫХ ИЗМЕРЕНИЙ ТЕМПЕРАТУРЫ НЕГЛУБОКОГО ВОДОЕМА

Л.А.Компаниец¹, О.С.Володько², Л.В.Гаврилова³

¹Институт вычислительного моделирования СО РАН, kla@icm.krasn.ru

²Институт океанологии АО РАН, olga.pitalskaya@gmail.com

³Сибирский федеральный университет, lv.gavrilova@gmail.com

Летом 2015 года Институтом Биофизики СО РАН на озере Шира проводились масштабные измерения температуры воды.

Измерения проводились в нескольких точках и в течение длительного периода.

В отличие от предыдущих лет, измерения проводились с использованием приборов, позволяющих мерить температуру с точностью до 0,1 С°.

В данной работе представлены результаты анализа данных измерений температуры воды за 2015 г. в озере Шира с использованием метода главных и независимых компонент.

Метод эмпирических ортогональных функций (метод главных компонент) хорошо зарекомендовал себя в задачах анализа данных течений в различных водоемах. В силу своей сложной пространственно-временной структуры эти течения представляют собой идеальный объект для применения метода главных компонент.

В статьях [1,2-6] приводится теоретическое обоснование метода главных компонент и результаты применения этого метода для анализа распределения тропических волн и анализа глубоководных и поверхностных течений в Северной Каролине.

Анализ независимых компонент [7-10] связан с проблемой разделения смесей сигналов с априори неизвестными характеристиками и в настоящее время широко используется в медицине и биологии, обработке изображений.

Для анализа температуры в водоеме он может быть использован непосредственно как способ устранения шума или выделения артефактов так, как он используется в ряде работ, например, в [11].

Озеро Шира – неглубокое сильно стратифицированное озеро с максимальной глубиной 25 м и температура воды мерилась до самого дна.

В результате измерений имеем таблицу с указанием места измерения, глубиной, на которой проводились измерения, и собственно данными по температуре.

N	E	Date	Time	0,00	0,48	1,42	2,37	2,84	3,30	3,77	4,23	4,70
54,522179	90,211041	30.05.2015	12:39:00	12,8	12,1	11,8	11,7	11,8	11,6	11,8	11,6	11,6
		30.05.2015	12:39:30	12,7	12	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:40:00	12,9	12,1	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:40:30	12,9	12	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:41:00	12,8	12	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:41:30	12,8	12	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:42:00	12,8	12	11,7	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:42:30	12,9	12	11,7	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:43:00	12,9	12	11,6	11,6	11,7	11,6	11,7	11,6	11,6
		30.05.2015	12:43:30	12,9	12,2	11,7	11,6	11,7	11,6	11,7	11,6	11,6

Рис.1. Фрагмент таблицы распределения температуры по глубине в точке Т3.

Следующий рисунок иллюстрирует положение точек измерения.

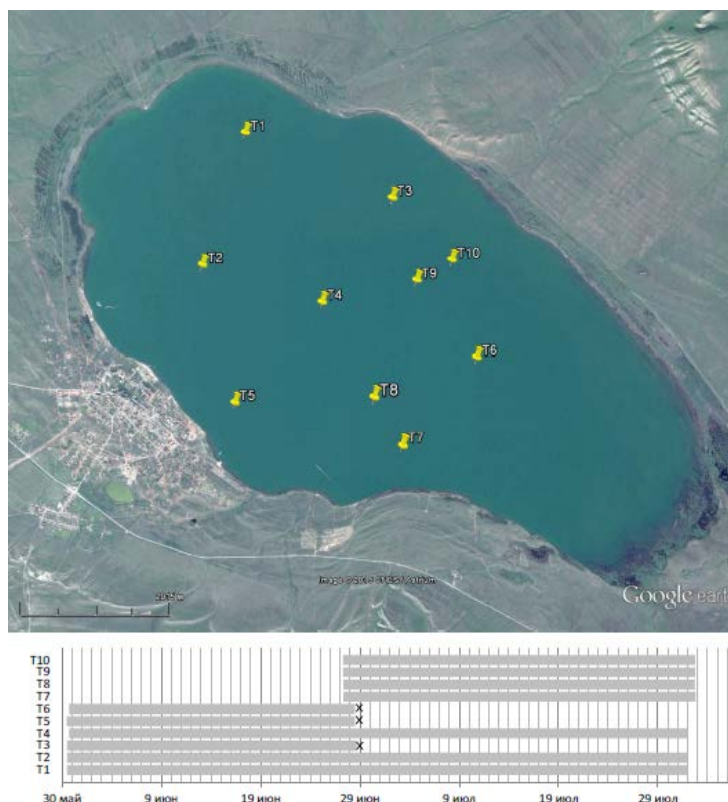


Рис.2. Расположение станций и периоды регистрации вертикальных профилей температуры в озере Шира в 2015 г. Рисунок предоставлен с.н.с. Института биофизики СО РАН, к.б.н. Толмеевым А.П.

Были проанализированы массивы температур в летний период в точках T1, T3, T4, T6.

Для выделения общей составляющей течения был применен метод главных компонент.

Цель этого метода – оптимально описать и проанализировать поле температур течений $T(z, t)$, где z – глубина бассейна.

Поле температур записывается в виде

$$T(z_i, t_k),$$

где $t_k (k = 1, \dots, K)$ и $z_i (i = 1, \dots, N)$ – переменные по времени и пространству соответственно.

В результате применения этого метода поле температур представляется в виде

$$T(z_i, t_k) = \sum_{n=1}^N E_n(t_k) \phi_n(z_i),$$

Функции ϕ_n – эмпирические ортогональные моды, E_n – соответствующие модальные коэффициенты.

Ниже на рис. 3-5 изображены первые 3 моды в точках T1, T3, T4, T6 и модальные коэффициенты в точке T4 (рис. 6).

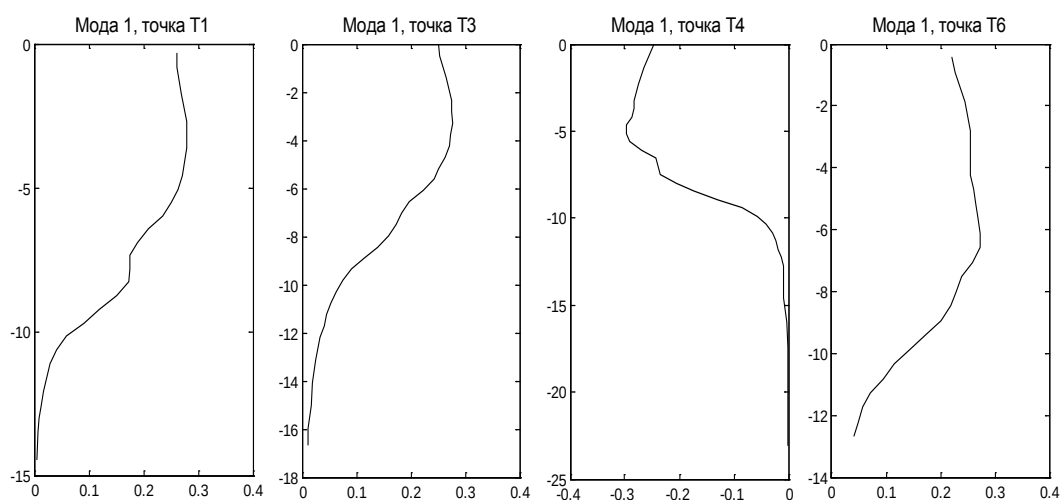


Рис. 3. Первая мода для T1, T3, T4, T6 (данные за период 30/05/15–28/06/15)

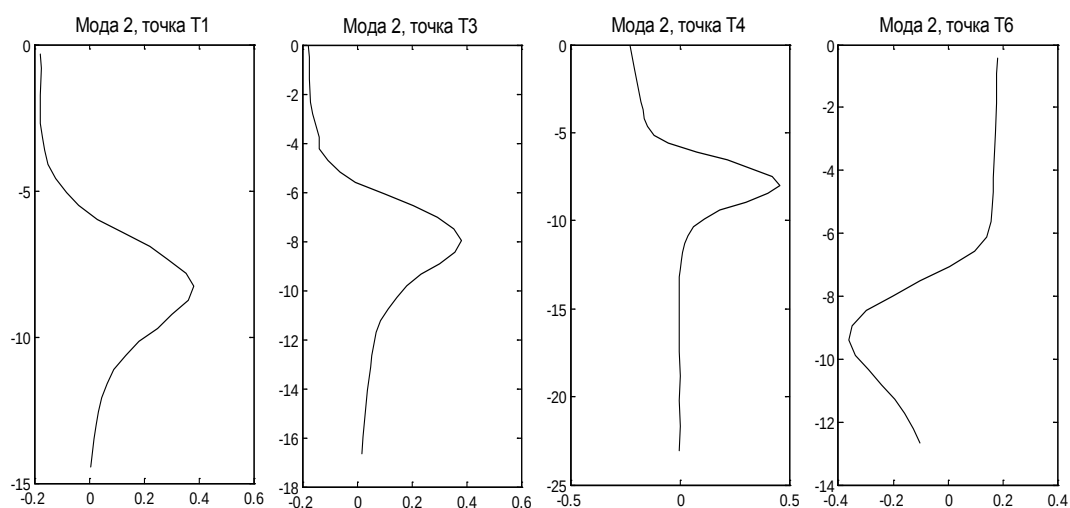


Рис. 4. Вторая мода для T1, T3, T4, T6 (данные за период 30/05/15–28/06/15)

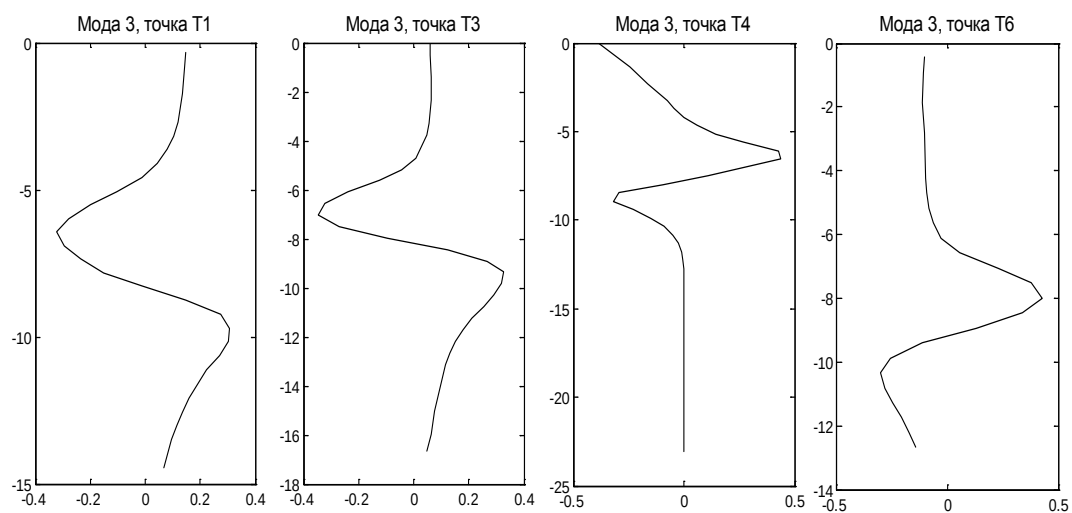


Рис. 5. Третья мода для T1, T3, T4, T6 (данные за период 30/05/15–28/06/15)

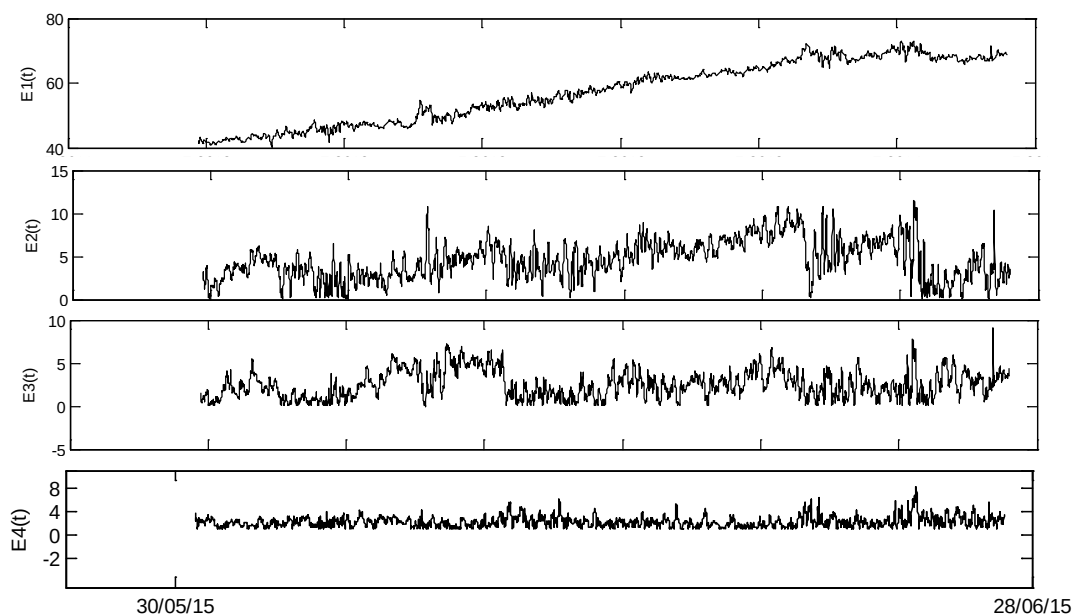


Рис.6. Модальные коэффициенты E_1, E_2, E_3, E_4 для измерения в точке Т4.

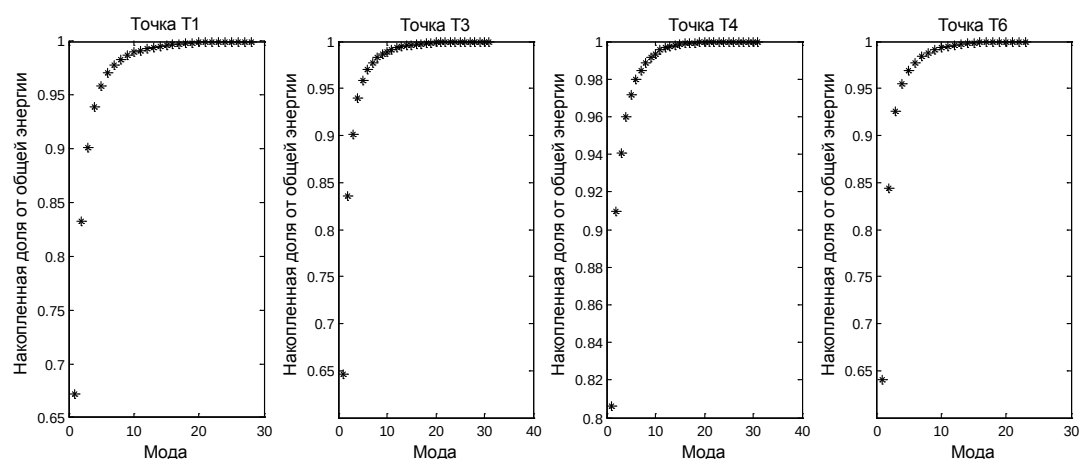


Рис.7. Накопленная энергия для Т1, Т3, Т4, Т6 (данные за период 30/05/15–28/06/15).

Во всех измерениях (рис. 7) для содержательного анализа достаточно первых восьми мод.

Первая мода (рис.3) не меняет знака по всей глубине. Поскольку каждая мода определяется с точностью до знака, то во всех случаях ее можно трактовать как температуру при наличии зоны термоклина. Влияние первой моды в глубоководной точке Т4 практически линейно увеличивается до двадцатого июня и далее не меняется до конца месяца (рис.6).

Вторая мода (рис.4) во всех измерениях имеет одну переменную знака и максимум значений, который приходится на зону термоклина.

С увеличением номера моды увеличивается количество перемен знака (рис. 5) аналогично тому, как это происходит в так называемых динамических модах, возникающих при исследовании внутренних волн [12]. Высокочастотные осцилляции, порождаемые модами с большими номерами, могут быть отфильтрованы в соответствии с рис. 4.

Также для указанного измерения Т1 было проведено разделение на независимые компоненты в соответствии со стандартным алгоритмом

$$X = WS,$$

где X – матрица данных исходного пространства, S – матрица независимых компонент, W – матрица весов для перехода из пространства S в пространство X .

Вычисления проводились с использованием алгоритма FastICA [8].

Анализируя независимые компоненты, можно выделить артефакты в данных (резкие скачки значений температуры на некоторых временных промежутках). В частности, на рис. 8 обозначен артефакт, вызванный возможными большими коротковолновыми осцилляциями на глубине 3 м (в перемешанном слое), которые требуют дальнейшего изучения (рис. 9).

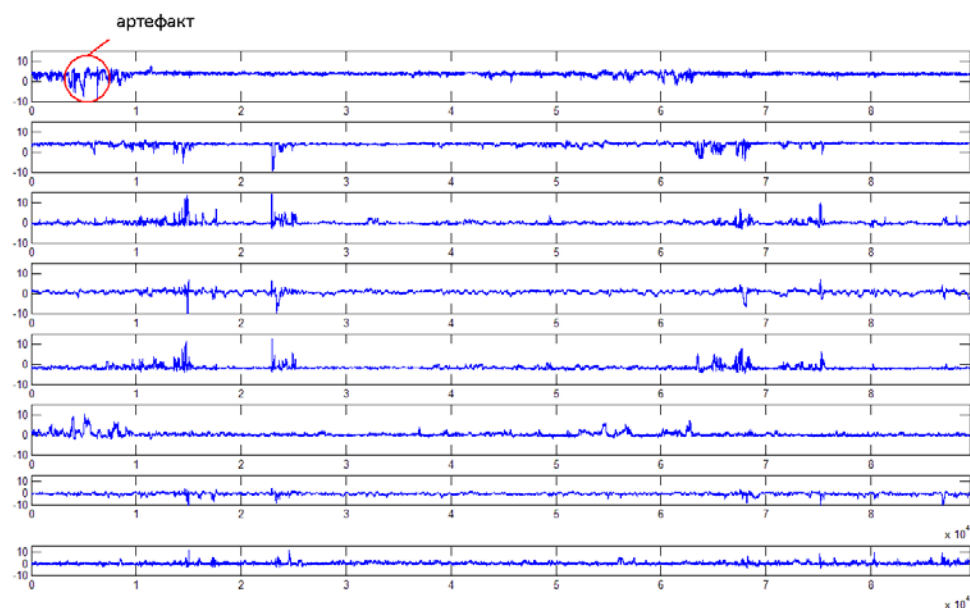


Рис. 8. Первые восемь независимых компонент для точки T1 (данные за период 01/07/15–01/08/15).

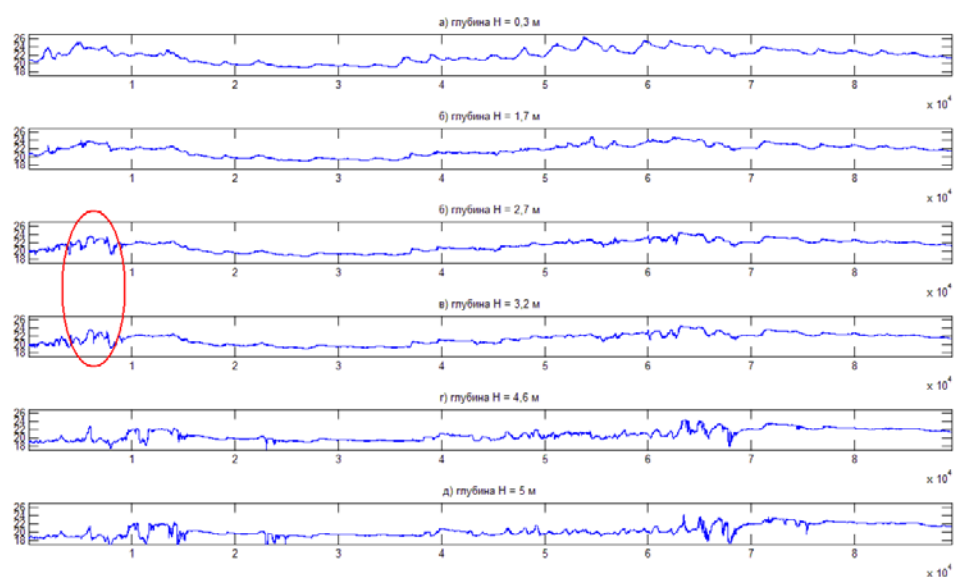


Рис.9 а) - д). Изменение температуры по времени на соответствующих горизонтах для точки T1 (данные за период 01/07/15–01/08/15).

Процедура удаления коротковолновых возмущений и артефактов является необходимой частью подготовки данных для уточнения частоты Брента-Вайсяля, используемой при исследовании частот внутренних волн.

Список литературы

1. Lorenz E.N. Empirical orthogonal functions and statistical weather prediction / E. Lorenz — Technical Report, Statistical Forecasting Project, Department of Meteorology, Massachusetts Institute of Technology, Cambridge, MA, 1956 — 49 p.
2. Wallace J. M. Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical consideration / J. M. Wallace, R.E. Dickinson // *Journal of Applied Meteorology*. — 1972 — Vol.11 — № 6 — P. 887-892.
3. Wallace J. M. Empirical orthogonal representation of time series in the frequency domain. Part II: Application to the study of tropical wave disturbances./ J. M. Wallace, R.E. Dickinson // *Journal of Applied Meteorology*. — 1972 — Vol.11 — № 6 — P. 893-900.
4. Kundu P.K. Some three-dimensional characteristic of low-frequency currents near the Oregon Coast / P.K. Kundu, J.S. Allen // *Journal of Physical Oceanography*. — 1976 — Vol.6 — P. 181-199.
5. Matsuura H. Wind Effects on Sub-Tidal Currents in Puget Sound/ H. Matsuura, A. C. Glenn // *Journal of Oceanography*. — 1997 — Vol.53 — P. 53-66.
6. Kaihatu J.M. Empirical Orthogonal Function analysis of ocean surface currents using complex and real-vector methods / J.M. Kaihatu, R. A. Handler, G. O. Marmorino, L. K. Shay // *Journal of atmospheric and oceanic technology*. — 1998 — Vol.15 — P.927-941.
7. Principal Manifolds for Data Visualisation and Dimension Reduction. Series: Lecture Notes in Computational Science and Engineering / Gorban A.N. [et al.]. — Berlin-Heidelberg-New York: Springer, 2007. - XXIV. - 340 p.
8. Hyvärinen A. A Fast Fixed-Point Algorithm for Independent Component Analysis / , Hyvärinen A., Oja E. // *Neural Computation*. — 1997 — 9(7) — P. 1483-1492.
9. Горячкин А.В. Метод анализа независимых компонент на основе преобразования независимости / Горячкин А.В., Шатских С.Я. // *Доклады Академии наук*. — 2004 — 398:4 — С. 466-468.
10. Ведмедь А.Г. Восстановление изображений с использованием анализа главных и независимых компонент / Ведмедь А.Г., Машталир С.В., Сакало Е.С. // *Системи обробки інформації*. — 2010 — выпуск 6 (87) — С. 66-72.
11. Захаров И.С. Компонентный анализ в задаче слепого разделения спонтанной электроэнцефалограммы / И.С. Захаров // *Электронный журнал «Молодежный научно-технический вестник*. — 2013 — №9.
12. Waves and oscillations in the ocean and in lakes. In *Continuum mechanics in environmental science and geophysics* / Edited by K.Hutter – Springer-Verlag Berlin, New York, 1993. – 434 p.

ПОСТРОЕНИЕ МОДЕЛИ ОЦЕНКИ И ПРОГНОЗА ПОСЕЩАЕМОСТИ ПРЕДПРИЯТИЙ ОБЩЕСТВЕННОГО ПИТАНИЯ НА ОСНОВЕ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ

Е.В. Попова¹, М.В. Гунер¹, О.И. Пятковский¹

¹Алтайский государственный технический университет им. И.И. Ползунова,
каф. информационных систем в экономике, horyzont1@mail.ru

В настоящее время отрасль общественного питания бурно развивается: растёт число заведений, улучшается качество обслуживания, расширяется ассортимент предлагаемых услуг. Согласно данным Росстата и оценке РБК.research, на российском рынке сейчас работают свыше 76 тыс. ресторанов, кафе и баров, для сравнения в 2010 г. их было лишь 60,5 тыс. [7].

Время как ресурс становится все более и более востребованным, в связи с чем питание в ресторанах и кафе, покупка кулинарной продукции, приготовленной профессионалами, кажутся потребителю более выгодными, чем трата времени на самостоятельное приготовление пищи.

Цель работы – построить модель оценки и прогноза посещаемости предприятий общественного питания на основе применения нейросетевых технологий.

В качестве входных факторов в модели были выбраны категория заведения и средний чек; географическое положение: расстояние до ближайшей остановки, количество маршрутов общественного транспорта, количество новостроек, университетов и других образовательных учреждений, бизнес-центров, других точек общепита и продуктовых магазинов в радиусе 500м; отзывы гостей заведений общественного питания, размещенные в открытом доступе; оценки экспертной группы: цена, вкуснота блюд, чистота зала, качество обслуживания и доброжелательность персонала.

Выбор нейронных сетей для решения поставленной задачи обусловлен их отличными аппроксимирующими способностями, возможностью извлечения знаний из массивов данных [1, 2].

С целью настройки модели оценки и прогноза, проведения вычислительных экспериментов, апробации модели на реальных данных и доказательства её адекватности нами была подготовлена выборка, содержащая сведения о более чем 60 предприятий общественного питания г. Барнаула.

Фрагмент выборки представлен в таблицах 1-2.

Таблица 1

Фрагмент выборки с данными о предприятиях общественного питания г. Барнаула.

Показатели оценки географического положения заведений

Наименование предприятия общественного питания (ресторана, кафе, столовой)	Категория	Оценка географического положения заведения (на основе данных 2ГИС, Yandex)							
		Расстояние до ближайшей остановки, м	Кол-во маршрутов общественного транспорта	Кол-во новостроек	Кол-во университетов, колледжей	Кол-во других учебных заведений	Кол-во офисов (бизнес-центр)	Кол-во других точек общепита	Кол-во продуктовых магазинов
Бочкари	Ресторан	120	24	2	3	7	-	29	7
Ипонкин brothers (Ленина)	Ресторан	280	22	-	7	5	2	8	1
Кинза Мята (Пионер)	Гриль-бар	150	25	1	1	1	-	18	3

Перцы & Ипонкин	Ресторан	350	14	3	1	4	-	12	5
Пора покушать корпус В	Столовая	290	21	-	5	1	1	12	4
Шашлыкофф (Балтийская)	Гриль-бар	70	8	19	-	4	3	8	5
Bistro (Димитрова)	Кафе	360	21	-	4	3	1	19	3

Таблица 2

Фрагмент выборки с данными о предприятиях общественного питания г. Барнаула.
Данные о заведениях из открытых источников

Наименование предприятия общественного питания (ресторана, кафе, столовой)	Данные о заведении из открытых источников			Данные собственных исследований						
	Средний чек, руб.	Отзывы (фламп)	Отзывы (ревизона)	Цена/качество	Вкус	Персонал	Чистота	Наличие доставки	Кол-во мест	Посещаемость (на момент посещения заведения экспертной группы), чел.
				оценки экспертной группы (по шкале от 2 до 5)						
Бочкари	900	3,9	3,6	4	5	4	5	-	150	43
Ипонкин brothers (Ленина)	680	3,6	-	5	5	5	4	+	120	24
Кинза Мята (Пионер)	800	3,8	-	4	5	3	5	+	200	50
Перцы & Ипонкин	580	3,2	3,8	5	4	4	4	+	200	20
Пора покушать корпус В	150	-	-	5	3	4	5	-	120	13
Шашлыкофф (Балтийская)	500	3,2	3,9	5	5	3	5	+	150	26
Bistro (Димитрова)	120	3,2	4,3	4	4	5	5	-	70	15

В качестве метода обучения нейронных сетей был выбран антиовражный метод kParTan [3]. Наилучший результат с ошибкой обобщения 9,509% был достигнут при использовании многослойного персептрона со структурой 4:2:1 (см. рис. 1).

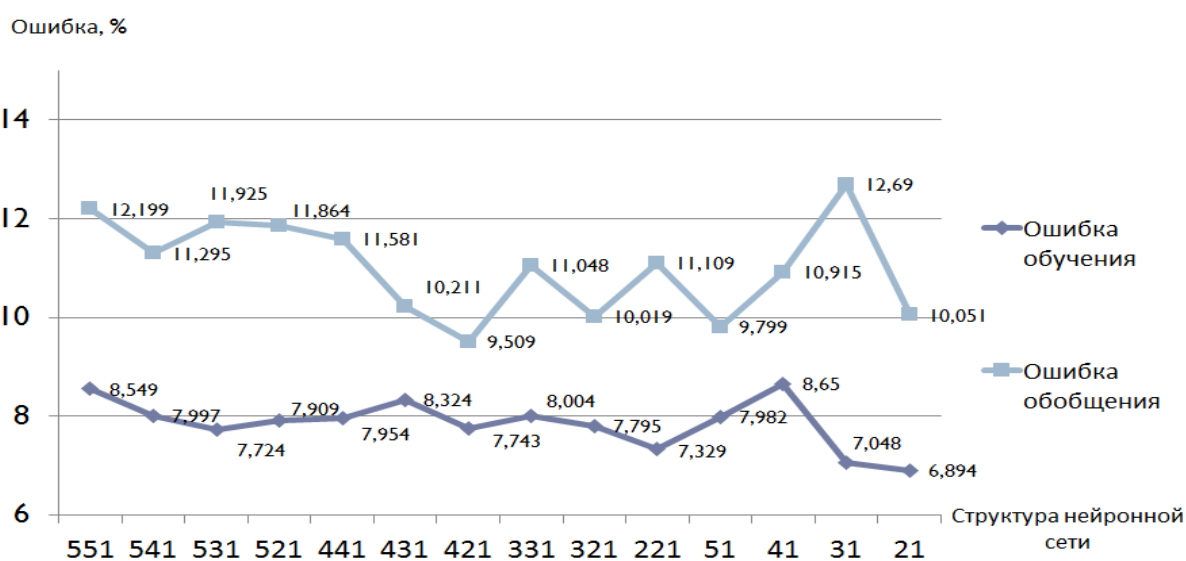


Рис. 1. Результаты экспериментов по обучению искусственных нейронных сетей различной структуры

Довольно интересным являются результаты анализа значимости входных параметров. Сравнивая ошибки обобщения, полученные после удаления соответствующих входов, с найденной ранее минимальной ошибкой обобщения, мы пришли к выводу, что самыми значимыми при оценке посещаемости предприятия общественного питания являются средний чек, расстояние до ближайшей остановки, количество бизнес-центров и других точек общепита в радиусе 500 м (см. рис. 2).

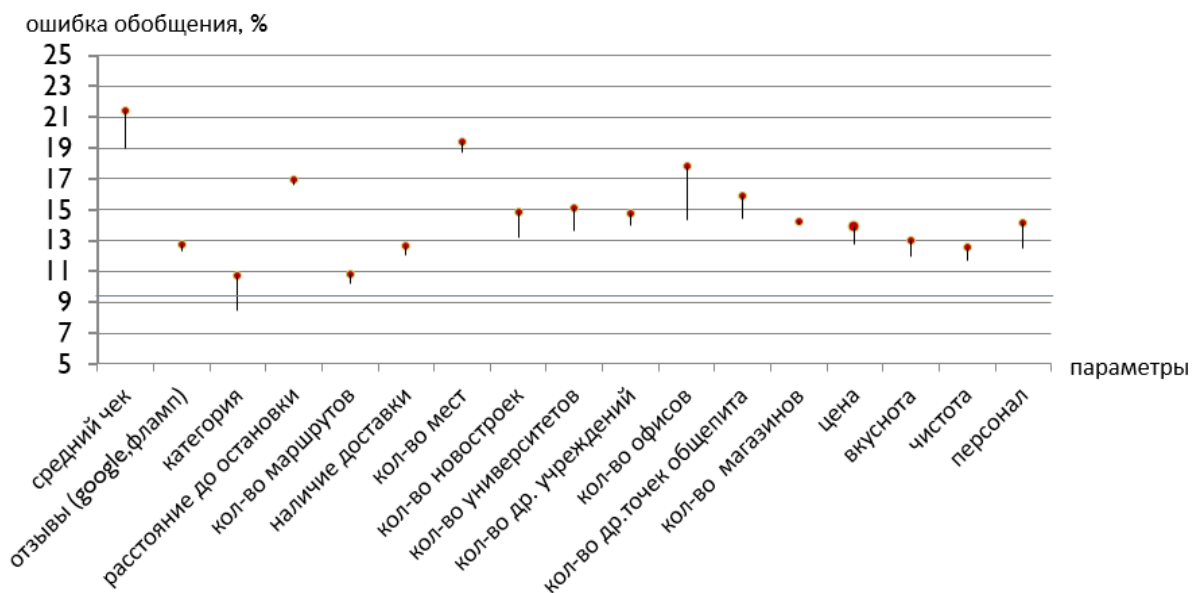


Рис. 2. Анализ значимости входных факторов. Распределение ошибок обобщения, полученных после удаления из выборки соответствующих входов

Также в ходе работы мы провели эксперименты, в которых уже обученную на реальных данных нейронную сеть использовали для решения задачи прогноза посещаемости учреждений общественного питания в зависимости от значений конкретных входных параметров.

Результаты экспериментов приведены на рис. 3-6.

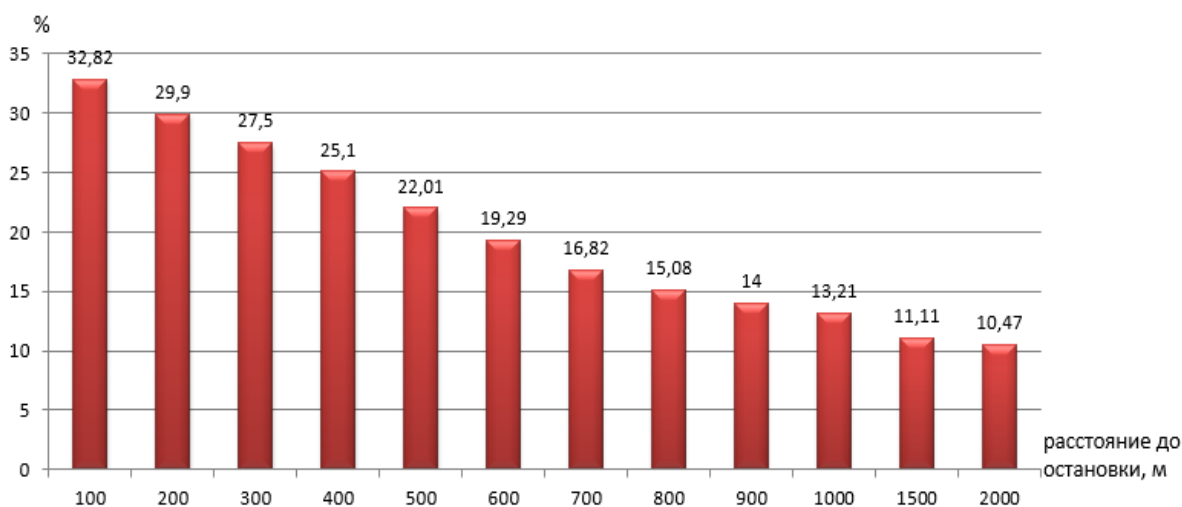


Рис. 3. Оценка средней посещаемости предприятий общественного питания (в процентах) в зависимости от расстояния до ближайшей остановки

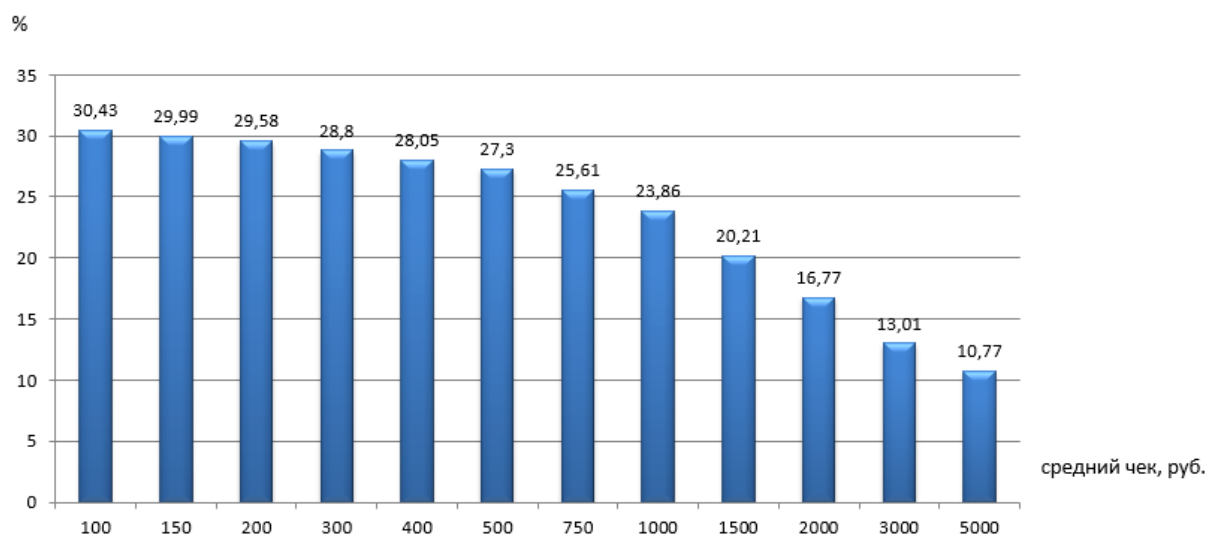


Рис. 4. Оценка средней посещаемости предприятий общественного питания (в процентах) в зависимости от среднего чека

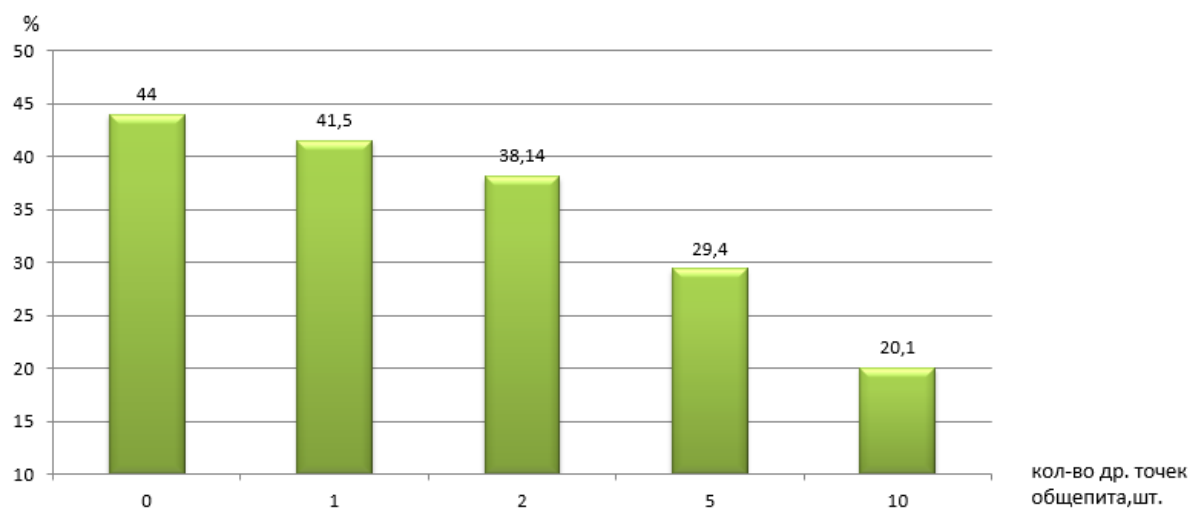


Рис. 5. Оценка средней посещаемости предприятий общественного питания (в процентах) в зависимости от количества других точек общепита в радиусе 500 м

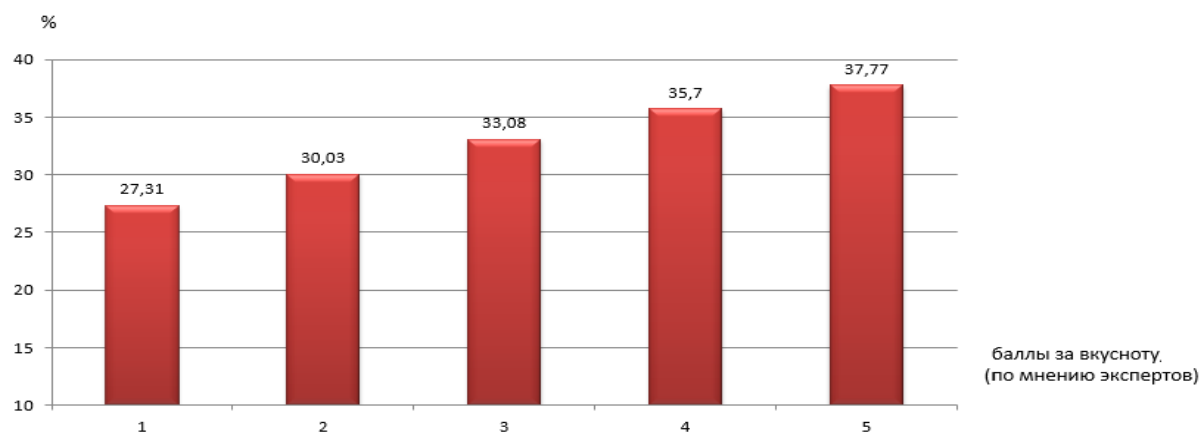


Рис. 6. Оценка средней посещаемости предприятий общественного питания (в процентах) в зависимости от вкусноты блюд (по мнению экспертов)

Методика оценки и прогноза посещаемости предприятий общепита была настроена в интеллектуальной информационной системе «Бизнес Аналитик», являющейся разработкой кафедры информационных систем в экономике ФГБОУ ВО «Алтайский государственный технический университет им. И.И. Ползунова» (см. рис. 7) [4, 5].



Рис. 7. Структура интеллектуальной информационной системы «Бизнес Аналитик»

С целью автоматизации процессов расчета по предложенной методике была создана прикладная информационная система управлениями учреждениями общественного питания, среда разработки «1С: Предприятие» версии 8.3. Интеграция новой системы с аналитическим комплексом «Бизнес Аналитик» осуществлялась посредством его облачных сервисов (см. рис. 8).



Рис. 8. Схема взаимодействия прикладной информационной системы управления учреждениями общественного питания с WEB-сервисами системы «Бизнес Аналитик»

Облачные сервисы «Бизнес Аналитика» реализованы в среде Microsoft Visual Studio на языке программирования C# с использованием технологии ASP.NET и доступны в интернете по адресу http://ba40.ru/ba_xml.aspx [6].

На основе результатов расчетов по нескольким десяткам точек общественного питания г. Барнаула мы построили карту, на которой цветом отметили наиболее и наименее посещаемые рестораны согласно нашему прогнозу (см. рис. 9). Подобная информация может быть очень полезна руководителю сети предприятий общественного питания при принятии решения об открытии новой точки.

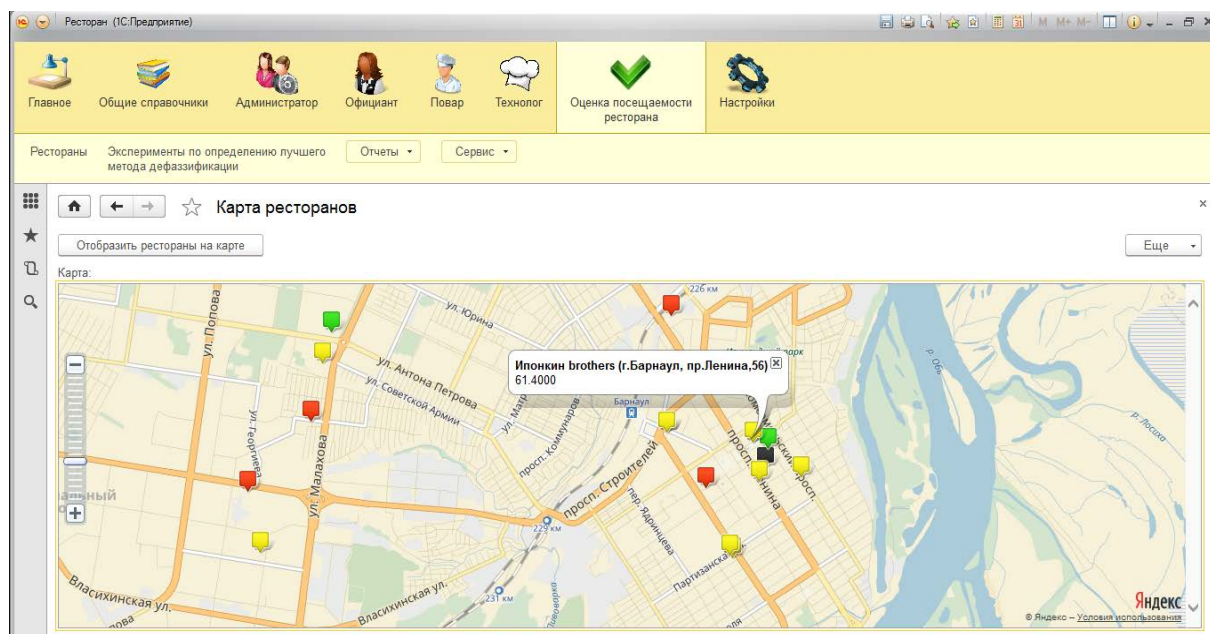


Рис. 9. Карта предприятий общественного питания г. Барнаула с оценками среднего уровня их посещаемости

В результате проведения настоящего исследования мы подтвердили особую важность географического положения для успешности заведений общественного питания, определили значимость конкретных факторов, влияющих на посещаемость.

Список литературы

1. Горбань А.Н. Обучение нейронных сетей / А.Н. Горбань. - М.: изд. СССР-США СП "ParaGraph", 1990. - 159с.
2. Матвеев, М.Г. Модели и методы искусственного интеллекта. Применение в экономике: учебное пособие / М.Г. Матвеев, А.С. Свиридов, Н.А. Алейникова. - М.: Финансы и статистика, ИНФРА-М, 2008. - 448с.: ил.
3. Миркес Е.М. Нейрокомпьютер. Проект стандарта / Е.М. Миркес. - Новосибирск: Наука. Сибирское предприятие РАН, 1999. - 337с.
4. Пятковский О.И. Интеллектуальные компоненты автоматизированных информационных систем управления предприятием: Монография / О.И. Пятковский. - Барнаул: изд-во АлтГТУ, 1999. - 351с.
5. Пятковский, О.И. Комплекс построения гибридных интеллектуальных систем «Бизнес аналитик» / О.И. Пятковский, М.В. Гунер // Ползуновский альманах №2. Выпуск посвящен Четвертой международной научно-практической конференции ВИС – 2009, 20 ноября 2009 г., АлтГТУ им. И.И. Ползунова, г. Барнаул, 2009. – с. 199-200
6. Пятковский, О.И. Разработка WEB-сервисов интеллектуальной информационной системы «Бизнес Аналитик» и их применение при решении задачи оценки инвестиционных проектов / О.И. Пятковский, М.В. Гунер, А.С. Авдеев // Нейроинформатика, ее приложения и анализ данных: Материалы XXI Всероссийского семинара, 27-29 сентября 2013 г. / под ред. А.Н. Горбаня; отв. за вып. М.Г. Садовский; - Красноярск; Институт вычислительного моделирования СО РАН, 2013, с. 111-117
7. Российский сетевой рынок общественного питания 2016. – М.: РБК, 2016. – 60 с. [Электронный ресурс]. - Режим доступа: http://alfabank.rbc.ru/media/research/file/Общепит_спецверсия.pdf

ПРОГНОЗИРОВАНИЕ ОЖИДАЕМОЙ ПРОДОЛЖИТЕЛЬНОСТИ ЖИЗНИ НАСЕЛЕНИЯ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ

Е.Н. Потылицына, Е.В. Сугак

Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева, *Leonova_en@mail.ru*

Средняя ожидаемая продолжительность предстоящей жизни (ОПЖ) - это число лет, которое проживет один человек в среднем из данного поколения родившихся при условии, что на всем протяжении жизни этого поколения смертность в каждой возрастной группе будет оставаться неизменной на уровне расчетного периода. Продолжительность предстоящей жизни рассчитывается для новорожденных (или иначе говорят — ожидаемая продолжительность жизни при рождении) и для достигших некоторого возраста «х» [1].

Ожидаемая продолжительность жизни (ОПЖ) - один из ключевых показателей, входящих в расчет индекса развития человеческого потенциала - главного интегрального показателя для межстрановых сравнений уровня и качества жизни населения. Кроме того, ОПЖ населения страны и отдельных ее регионов характеризует уровень жизни людей. Изменение ожидаемой продолжительности жизни может быть следствием проведения экономических, политических, социальных реформ, а также изменением идеологии общества [2].

ОПЖ при рождении служит признанным индикатором качества жизни и здоровья населения [3,4]. Прогноз этого показателя является весомым элементом в управлении общественными процессами, он необходим для анализа перспектив развития различных сфер, включая экологическую обстановку в отдельно взятом исследуемом регионе. Прогнозируя ожидаемую продолжительность жизни, важно не только определить будущие характеристики населения, но и предвосхитить негативные тенденции, деформирующие демографические процессы, проанализировать их предпосылки. Знание перспектив развития общества поможет целенаправленно воздействовать на социальноэкономические явления, скорректировать их, ликвидировать диспропорции.

По мнению экспертов ВОЗ, данный показатель является одним из показателей общего качества жизни в стране и в отдельно взятом регионе.

Согласно данным Территориального органа Федеральной службы государственной статистики по Красноярскому краю [5-8], ожидаемая продолжительность жизни в Красноярском крае стабильно растёт, хотя и наблюдается резкая диспропорция между ожидаемой продолжительностью жизни мужчин и женщин, что является крайне негативным показателем состояния любого социума [9].

В настоящее время имеется большой арсенал методов для прогнозирования и выявления эффекта различных воздействий окружающей среды на состояние здоровья жителей, в том числе и на ОПЖ. В целях оценки влияния различных факторов (экологических, социальных, экономических) используются различные методы.

Для прогнозирования исследуемого показателя был выбран современный метод интеллектуального анализа данных – искусственные нейронные сети (ИНС).

Этот метод находит применение в тех ситуациях, когда обычные методы анализа трудно или невозможно применить из-за отсутствия сведений о характере или закономерностях исследуемых процессов, взаимозависимостях явлений, фактов, о поведении объектов и систем из различных предметных областей, в том числе – в социальной и экономической [10].

Для построения нейросетевой модели использовалась свободно распространяемая программа NeuroPro [11]. Программа NeuroPro является свободно распространяемой альфа-версией нейросетевого программного продукта для работы с искусственными нейронными сетями и извлечения знаний из таблиц данных с помощью нейронных сетей в среде Windows. Разработчик - В. Г. Царегородцев, Институт вычислительного моделирования СО РАН г. Красноярск.

К возможностям программы относятся:

- работа (чтение, запись, редактирование) с файлами данных, представленными в форматах dbf (СУБД dBase, FoxPro, Clipper) и db (СУБД Paradox)
- создание слоистых нейронных сетей для решения задач прогнозирования (число слоев нейронов - до 10, число нейронов в слое - до 100);
- обученная нейронная сеть может одновременно решать несколько задач прогнозирования; для каждого из выходных сигналов могут быть установлены свои требования к точности прогнозирования.
- обучение нейронной сети производится по принципу двойственного функционирования с применением одного из следующих методов оптимизации;
- вычисление и вывод значимости входных сигналов сети (это оценка влияния входных сигналов на изменения выходных сигналов, полученная в первом приближении (т.е. с использованием только первых производных разложения в ряд Тейлора); и др.

Применение данного программного продукта возможно в традиционных областях, а именно, в медицине, экологии, климатологии, метеорологии, при построении моделей технических объектов и их идентификации, в экономике (прогнозирование курсов валют, акций и т. д.) и вообще, для решения любой задачи классификации или прогноза, которая решается при наличии выборки данных и для решения которой ранее использовались традиционные математические методы (регрессионный анализ, непараметрическая статистика), однако не была достигнута требуемая точность прогноза.

Данный нейропакет несомненно удобен для изучения и использования по следующим причинам:

- является русскоязычным,
- прост в изучении,
- имеет такие достоинства, как возможности упрощения сети и выявление наиболее (или наименее) значимых входов.

К недостаткам можно отнести скромные интерфейсные и сервисные удобства [11].

В качестве входных параметров для обучения ИНС использовались концентрации загрязняющих веществ (диоксид азота, пыль, оксид углерода.) находящихся в воздухе Красноярского края [5-8]. В обучении ИНС также были учтены выбросы от стационарных источников (с учетом Норильска) (тыс. тонн в год) и выбросы от автотранспорта (с учетом Норильского промрайона) (тыс. тонн в год). Помимо этого, ИНС позволяет учесть и ряд социальных параметров. В работе использовались: среднедушевые денежные доходы населения (в месяц), рублей и средний размер назначенных месячных пенсий (руб.). В таблице 1 представлены данные, используемые для обучения ИНС.

«Выходом» при обучении ИНС использовались данные об ожидаемой продолжительности жизни (при рождении) жителей Красноярского края с 1995 по 2013 гг. [5-8].

Прогностические возможности полученной ИНС оценивали двумя способами: расчет относительной ошибки полученных значений взятых по модулю (в %); расчет

средней абсолютной ошибки в процентах (MAPE). MAPE подчеркивает насколько велика ошибка прогноза в сравнении с действительными значениями ряда [12].

Таблица 1

Параметры, используемые для обучения ИНС

Год	Выбросы от стационарных источников (с учетом Норильска) (тыс. тонн в год)	Выбросы от автотранспорта (с учетом Норильского промрайона) (тыс. тонн в год)	Диоксид азота	Твердые вещества	Оксид углерода	Среднедушевые денежные доходы населения (в месяц), рублей	Средний размер назначенных месячных пенсий, руб.	Ожидаемая продолжительность жизни при рождении (число лет)
2015	2475,9	253,2	90300	124200	226000	26853,7	12994,2	69,7
2014	2355,8	236,2	88900	112700	201600	24806,2	11707,4	69,2
2013	2497,3	313,0	94200	115600	243700	24921,7	10781,9	69,1
2012	2582,7	446,8	93900	140000	244000	22524,3	9840,8	68,4
2011	2516,8	428,0	98200	133900	199800	20145,5	8887,2	68,3
2010	2480,2	406,6	109040	138052	183239	18055,2	8145,2	67,6
2009	2433,3	382,0	82500	124300	135540	16661,0	6639,8	67,5
2008	2478,6	407,7	93200	146840	140180	15604,5	4896,0	66,9
2007	2509,6	345,2	78600	149990	157480	12654,1	3949,0	66,6
2006	2466,1	295,3	66900	148710	151010	9640,2	3044,2	65,6
2005	2521,4	175,5	80600	141300	159750	7709,9	3949,0	63,1
2004	2528,8	167,8	59655	136520	145201	6408,0	2165,7	63,6
2003	2462,3	176,8	61593	138805	120196	5509,0	1831,7	62,7
2002	2458,1	164,8	59849	134359	117941	4346,0	1531,5	63,0
2001	2562,6	207,9	59354	142307	118802	3572,0	1173,6	63,0
2000	2640,4	192,0	74262	175496	135022	2773,3	853,5	62,5
1999	2644,2	174,5	69565	169850	133637	1954,0	562,0	63,1
1998	2617,1	166,8	70315	164659	148784	1179,0	437,0	63,0
1997	2670,8	151,6	67500	175400	157400	1150,0	421,0	64,4
1996	2610,9	203,2	72950	184114	174925	932,0	320,6	63,5
1995	2505,9	261,4	66900	179400	143529	594,5	282,3	62,3

Результаты и обсуждение

Обучение нейронной сети проводилось на статистических данных в период с 1995 по 2013 гг. Данные за 2014-2015 гг. были использованы для проверки прогностических возможностей полученной ИНС. Полученная в ходе обучения ИНС состояла из 3 слоев, в каждом слое по 10 нейронов. Число входов сети – 7, выход – 1.

Средняя относительная ошибка прогноза взятая по модулю составила 0,49%, MAPE составила 5%, т.е. сеть обеспечивает хорошую сходимость расчетных и фактических значений, наглядно это представлено на рисунке и таблице 2.

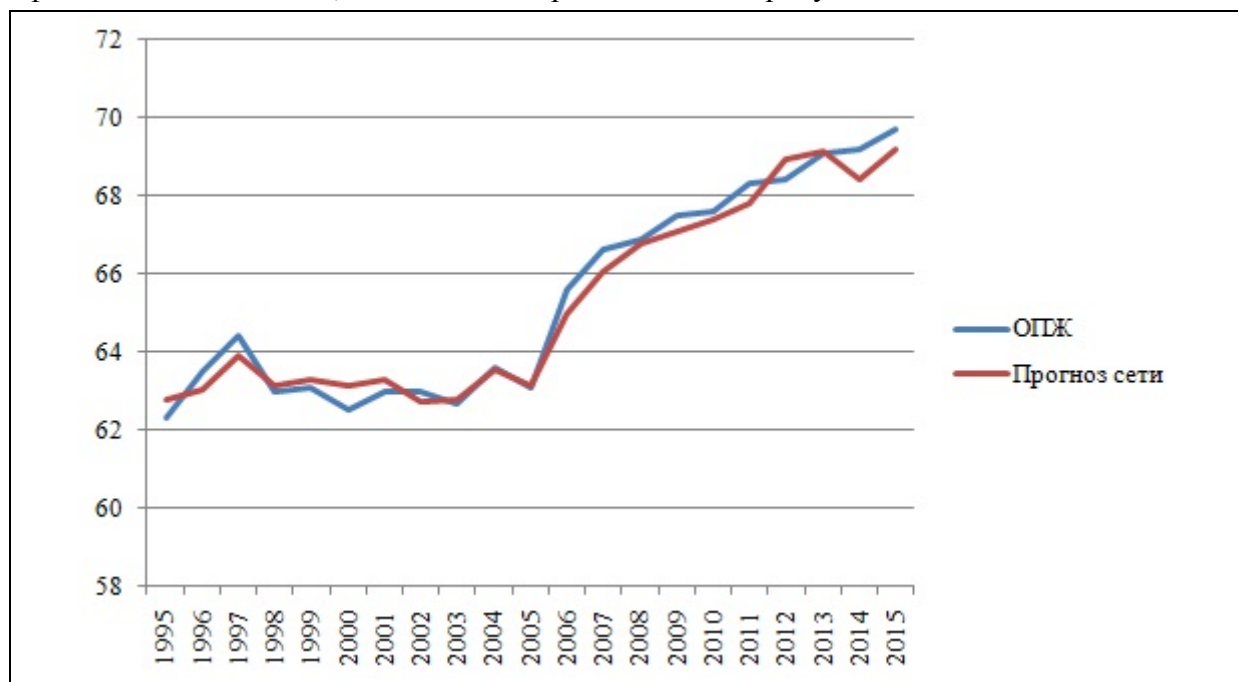


Рисунок. Графическое представление результатов обучения ИНС

Таблица 2

Расчет ошибки прогноза ИНС

Год	ОПЖ	Прогноз сети	Ошибка	Относительная ошибка, %	MAPE, %
2015	69,7	69,16	0,54	0,77	8,45
2014	69,2	68,44	0,76	1,10	12,03
2013	69,1	69,14	-0,04	0,053	0,57
2012	68,4	68,92	-0,52	0,76	8,18
2011	68,3	67,82	0,48	0,70	7,53
2010	67,6	67,37	0,23	0,34	3,62
2009	67,5	67,07	0,43	0,64	6,79
2008	66,9	66,78	0,12	0,19	1,97
2007	66,6	66,07	0,53	0,79	8,28

2006	65,6	65,00	0,61	0,94	9,68
2005	63,1	63,13	-0,03	0,05	0,50
2004	63,6	63,54	0,06	0,09	0,92
2003	62,7	62,79	-0,09	0,15	1,44
2002	63,0	62,75	0,25	0,40	3,94
2001	63,0	63,32	-0,32	0,50	4,99
2000	62,5	63,12	-0,62	0,99	9,72
1999	63,1	63,27	-0,17	0,27	2,66
1998	63,0	63,14	-0,14	0,22	2,22
1997	64,4	63,89	0,51	0,78	7,97
1996	63,5	63,05	0,45	0,71	7,14
1995	62,3	62,80	-0,50	0,81	7,94
Среднее значение (%)				0,49	5,06

Для проведения тестирования прогностических возможностей полученной ИНС были введены данные за 2014 и 2015 гг. Был получен прогноз значений ОПЖ за 2014 и 2015 гг. (Таб. 2) и рассчитаны средняя относительная ошибка (0,94%) и MAPE (10,24%).

При этом значения коэффициентов значимости входных параметров составили:

- выбросы от стационарных источников (с учетом Норильского промрайона) (тыс. тонн в год) – 1,00;
- выбросы от автотранспорта (с учетом Норильского промрайона) (тыс тонн в год) – 0,19;
- диоксид азота – 0,04;
- пыль – 0,34;
- оксид углерода – 0,19;
- среднедушевые денежные доходы населения (в месяц), рублей – 0,10;
- средний размер назначенных пенсий (в месяц), рублей – 0,80.

Таким образом, из полученных результатов следует, что в ОПЖ наиболее весомый вклад вносят выбросы от стационарных источников и социально-значимый показатель – размер ежемесячных пенсий.

Выводы

Полученная модель дает возможность по имеющимся базам данных социальных, экологических показателей и ожидаемой продолжительности жизни автоматически получать зависимости, отражающие причинно-следственные связи между зависимыми параметрами.

Исследуемая нейросетевая модель адаптирована к условиям региона и значительно упрощает и ускоряет оценку и прогноза развития социально-экологической обстановки в регионе в результате техногенного воздействия на окружающую среду и здоровье населения без масштабных клинических биомедицинских исследований [13]. Интерпретация полученных результатов позволяет оценить социально-экономическую эффективность планируемых технических, технологических, административно-

хозяйственных и управленческих решений в промышленном развитии региона и управлении охраной окружающей среды.

Полученные результаты свидетельствуют, что применение нейросетевых технологий для решения прикладных социально-экологических задач, связанных с обработкой информации и построением прогностических моделей, является перспективным направлением исследований.

Список литературы

1. Борисов В. А. Демография / А.В. Борисов. - М.: Издательский дом NOTABENE, 1999, 2001. - 272 с.
2. Звездина Н.В. Ожидаемая продолжительность жизни в России и факторы, влияющие на нее / Н.В. Звездина, Л.В. Иванова // Вопросы статистики. – 2015. - №17. – С.10-20.
3. Прохоров Б.Б. Условия жизни населения и общественное здоровье / Б.Б. Прохоров, И.В. Горшкова, Е.В. Тарасова // Проблемы прогнозирования. 2003. № 5. С. 127–139.
4. Щербакова Е.М. Евразийский демографический барометр / Е.М. Щербакова // Демоскоп Weekly: Электронная версия бюллетеня «Население и общество». 2007. № 285–286, 16–29 апреля (<http://www.demoscope.ru/weekly/2007/0285/barom01.php>)
5. Красноярский край в цифрах 2015: Стат.сб./ Красноярскстат. – Красноярск, 2016. – 105 с.
6. Красноярский край в цифрах 2012: Стат.сб./ Красноярскстат. – Красноярск, 2011. – 154 с.
7. Красноярский край в цифрах 2008: Стат.сб./ Красноярскстат. – Красноярск, 2009. – 138 с.
8. Красноярский краевой статистический ежегодник, 2016: Стат.сб./ Красноярскстат. – Красноярск, 2016. – 498 с.
9. Немировский В.Г. Социокультурный портрет Красноярского края : / монография / В.Г. Немировский, А.В. Немировская. – Красноярск : Сибирский юридический институт МВД России, 2010. – 263 с.
10. Сугак Е.В. Прикладная теория случайных процессов. Основные положения и инженерные приложения / Е.В. Сугак, Е.Н. Окладникова.- Красноярск: Сиб. гос. аэрокосмич. ун-т, 2006.- 168 с.
11. Круглов В.В. Искусственные нейронные сети. Теория и практика / В.В. Круглов, В.В. Борисов. - М.: Горячая линия-Телеком, 2002. - 382 с.
12. Ханк Д.Э. Бизнес - программирование, 7-е издание / Д.Э. Ханк, Д.У. Уичерн, А. Дж. Райтс. : Пер с англ. – М.: Издательский дом «Вильямс», 2003. – 656 с.
13. Потылицына Е.Н. Оценка социально-экологических рисков населения с использованием искусственных нейронных сетей / Е.Н. Потылицына, Е.В. Сугак // Экологический риск. Мат. IV Всероссийской научной конф. – Иркутск, 2017, с.325-327.

О НЕКОТОРЫХ МЕТОДАХ КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ, УЧИТЫВАЮЩИХ ТОПОЛОГИЮ ПРИБЛИЖАЮЩЕГО ИХ МНОГООБРАЗИЯ МАЛОЙ РАЗМЕРНОСТИ³

М.Г.Садовский^{1,2}, А.Н.Остыловский², Е.Ю.Бушмелев¹

¹Институт вычислительного моделирования СО РАН, msad@icm.krasn.ru

²Сибирский федеральный университет, hinayana@g-service.ru

Кластеризация является мощным и продуктивным методом анализа многомерных данных; неформально она заключается в выделении групп точек из большого ансамбля, в каждой из которых точки «ближе» друг к другу, чем к остальным. Очевидно, что результат такого выделения подмножества (либо подмножеств) точек существенно зависит от метода, которым они выделяются: вполне возможно, что исследователь сможет обнаружить кластерную структуру, существующую в смысле одного метода и не существующую в смысле другого. Дадим теперь строгие определения и приведём точные понятия. Пусть $\mathbf{R}^n$ — n -мерное метрическое пространство (всюду далее, кроме специально оговоренных случаев, мы будем подразумевать Евклидово пространство); величину n будем называть размерностью пространства. Далее, пусть $\mathbf{F}_M$ — множество (конечное) точек x_i , $1 \leq i \leq M$, $\forall i \ x_i \in \mathbf{R}^n$. Как говорилось выше, будем считать, что в $\mathbf{R}^n$ задана метрика, в частности, Евклидова метрика.

Задача кластеризации заключается в построении разбиения множества $\mathbf{F}_M$ на непересекающиеся подмножества (кластеры либо классы) $\mathbf{f}_{M_j}$, $1 \leq j \leq K$, так, чтобы

$$\bigcup_{j=1}^K \mathbf{f}_{M_j} = \mathbf{F}_M \quad \text{и при этом} \quad \forall l, l \neq j \ \mathbf{f}_{M_l} \cap \mathbf{f}_{M_j} = \emptyset.$$

Здесь индекс j перечисляет кластеры, K — общее число кластеров, M_j — число элементов (точек) в j -ом кластере. Забегая вперёд, сразу скажем, что очень часто кластеризация приводит к построению такого разбиения на подмножества, объединение которых меньше исходного множества: это означает, что некоторые элементы множества $\mathbf{F}_M$ не поддаются кластеризации и не могут быть надёжно включены ни в один кластер. Судьбу таких элементов следует изучать отдельно.

Разнообразие и число методов кластеризации столь велико, что их простое перечисление не может быть представлено в рамках настоящей статьи [1, 2]. Тем не менее, едва ли можно считать, что этот список замкнут и все мыслимые методы уже представлены. В рамках настоящей работы мы представим два новых метода анализа многомерных данных, один из которых базируется на исследовании топологических и метрических свойств соответствующего графа, а второй, строго говоря, не являясь методом кластеризации в узком смысле, направлен на определение топологии приближающего многообразия многомерных данных.

Начнём с изложения первого метода. Пусть имеется описанное выше множество $\mathbf{F}_M$ точек, и для каждой пары точек $x_i, x_j \in \mathbf{F}_M$ определено расстояние $\rho(x_i, x_j)$. Построим для множества $\mathbf{F}_M$ полносвязный граф G_M , в котором точки представляют собой вершины графа, а рёбра — расстояния между ними в выбранной метрике. Всего в таком графе найдётся $[M(M-1)]/2$ рёбер. Начнём последовательно исключать из этого графа G_M рёбра, начиная с самого длинного; как только граф потеряет связность, будем считать построение кластеризации оконченной.

³ Работа выполнена при поддержке гранта № 14.Y26.31.0004 Правительства Российской Федерации.

В описанной процедуре есть ряд «тонких мест», требующих пояснения. Первая проблема — проверка связности графа. Хорошо известно, что прямая проверка связности графа эквивалентна проверке разложимости матрицы [3]; проверка же разложимости является NP-полной задачей. К счастью, существуют эффективные альтернативные методы проверки связности. В нашем случае таким методом является поиск компонент связности графа; хорошо известно, что эта задача имеет сложность не выше $O(N^2)$, где N — число вершин графа G_M .

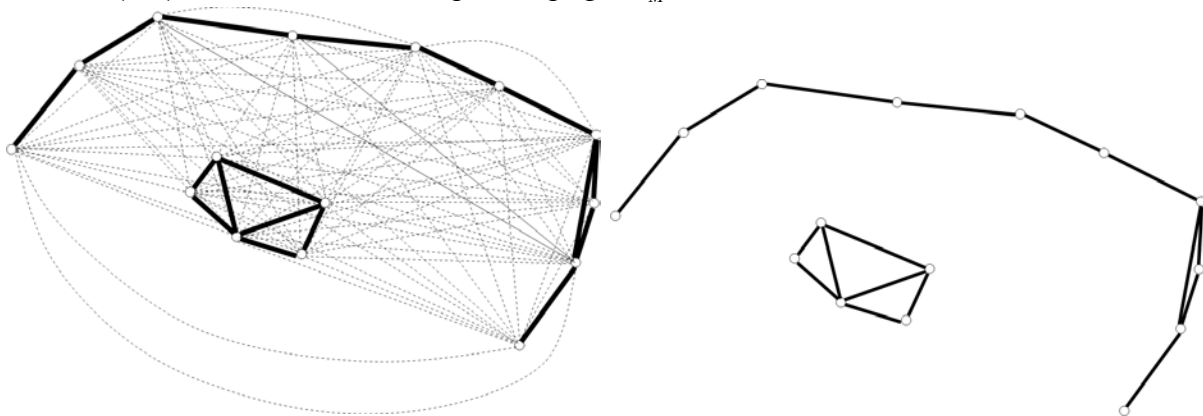


Рис.1. Пример построения кластеризации методом исключения наидлиннейших рёбер во взвешенном графе.

Наивная оценка сложности поиска компонент связности вида $O(N^3)$, подразумевающая, что их поиск следует проводить после **каждого** единичного исключения наидлиннейшего ребра, неверна. Действительно, рассмотрим полносвязный граф G_M с M вершинами; в худшем (то есть, при исключении минимального числа рёбер) случае он распадётся на две несвязные компоненты — тривиальный граф, состоящий из одной вершины и полносвязный подграф G_{M-1} с $M-1$ вершинами. Для этого потребуется исключить $M-1$ ребро. Тем самым, первую проверку на связность надо проводить после исключения $M-1$ ребра, вторую — $M-2$ рёбер и так далее, что делает задачу проверки связности графа с исключёнными рёбрами существенно менее затратной.

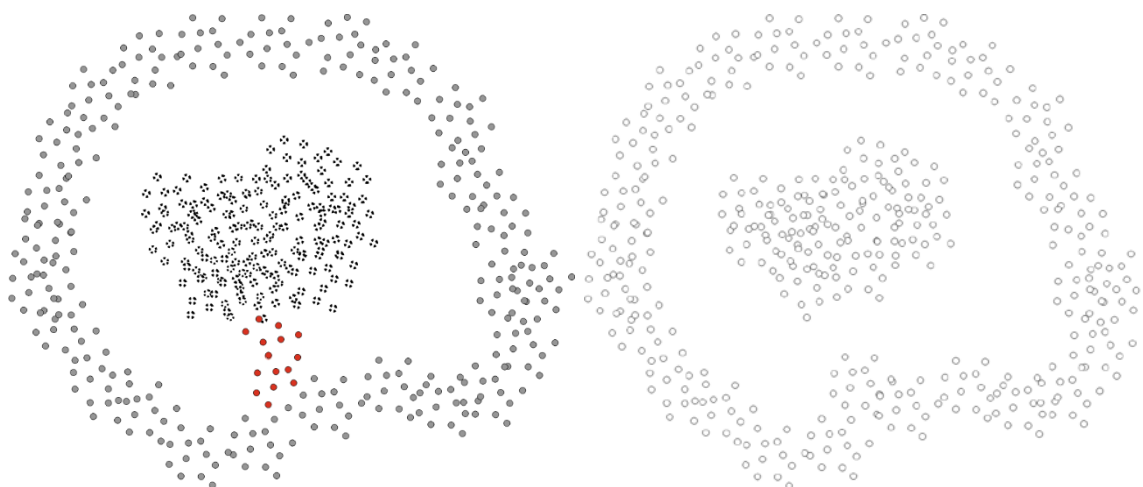


Рис.2. Пример кластеризации с «перемычками».

Пример построения такой кластеризации показан на рис.1. Следует отметить, что идея такого метода — исключать рёбра с экстремальным (не обязательно определяемым метрическим образом) весом — была впервые предложена в работе [4]. Однако авторы этой работы не учли одного важного обстоятельства, следующего из метрических свойств того графа, который может быть получен путём исключения наидлиннейших рёбер. В настоящей работе мы исправляем этот недочёт и представляем метод, позволяющий учитывать такие особенности.

Рассмотрим пример конфигурации, показанный на Рис.2; следует особо подчеркнуть, что на этом рисунке представлен двумерный случай, который может вводить в заблуждение своей наглядностью. Тем не менее, все изложенные ниже результаты остаются верными для метрического пространства произвольной конечной размерности. Итак, на Рис.2 показаны два кластера (справа), которые невозможно выделить описанным выше методом удаления наидлиннейших рёбер из (первоначально) полносвязного графа: мешает «перемычка», показанная в левой части рисунка красным цветом. Вместе с тем, не вызывает сомнения тот факт, что её удаление совершенно точно разобьёт всё множество на два кластера: «бублик» и шар, располагающийся внутри. Попутно заметим, что другие методы линейной кластеризации (метод динамических ядер и т.п.) также не способны выделить эти кластеры.

Для данных подобного типа предложен следующий метод кластеризации. На первом шаге построим граф, содержащий лишь минимальные расстояния (точнее, расстояния, меньшие определённого порога); это позволит выявить картину типа той, что представлена на Рис.2 слева. Следом встаёт задача выявления «перемычки». Идея поиска перемычки достаточно проста и прозрачна: перемычка — это такое подмножество $\Lambda \subset \mathbf{F}_M$ точек, которое а) весьма мало по сравнению со всем множеством $\mathbf{F}_M$ ($|\Lambda| \ll |\mathbf{F}_M|$, здесь $|\cdot|$ — мощность множества) и б) через вершины $f_j \in \Lambda$ проходит патологически много длинных путей в графе. Всё сказанное позволяет построить конструктивную процедуру поиска перемычек.

Будем считать, что методом исключения рёбер с наибольшим весом была получена конфигурация типа той, что показана на Рис.2 слева; пусть G_L^* — граф, соответствующий этой конфигурации, $L < M$. Для начала найдём в таком графе все ближайшие расстояния: эта задача не является трудоёмкой, поскольку эквивалентна построению минимальной компоненты связности на взвешенном графе [3], а общее число проверок расстояний не превышает $L/2$, где L — число вершин в графе G_L^* . Собственно идея поиска перемычки сводится к нахождению достаточно длинных путей в получившемся графе и вычислению частот встречаемости вершин в этих путях. Подчеркнём, что говоря здесь о расстояниях, мы имели в виду метрическое расстояние между точками в том пространстве, в котором они заданы, а теперь переходим к расстоянию на графах (числу рёбер, соединяющих нужные нам вершины).

Итак, L — число вершин в получившемся графе G_L^* . Общее число пар вершин равно $[L(L-1)]/2$, однако число путей может быть существенно больше: в целом индекс связности графа G_L^* остаётся весьма высоким. Оценку числа путей, которые необходимо построить для нахождения перемычки, можно получить по величине диаметра $d(G_L^*)$ графа G_L^* ; ещё более точную оценку можно получить, исходя из распределения расстояний (в смысле расстояний на графах), наблюдаемых на графе G_L^* .

Пусть величина d известна; тогда для нахождения перемычки потребуется найти достаточное количество n_i путей длины $0,5d < s \leq d$. Грубая оценка числа n_i таких путей даётся соотношением

$$n_{\alpha} = \frac{n(n-1)}{4}, \quad (1)$$

где n — число вершин в графе G_L^* . Эта оценка подразумевает, что приблизительно половина путей в графе G_L^* имеют достаточно большую длину. Однако оценка (1) является весьма завышенной: достаточно рассмотреть n путей в графе G_L^* . Рассмотрим набор из $\sqrt{n}$ случайно выбранных вершин в графе G_L^* ; тогда общее число путей между ними будет близко к n . Осталось построить все эти пути и выбрать среди них такие, длина l которых $l \approx d/2$ (здесь d — диаметр графа).

На следующем шаге нужно составить список всех вершин графа G_L^* , вошедших в эти пути и подсчитать, сколько раз они встречаются в упомянутых путях. Вершины x_i^* , такие, что а) $i \ll n$ и б) $k(x_i^*) \approx n/2$ и составляют перемычку; здесь по-прежнему n — число вершин в графе G_L^* , а $k(x_i^*)$ — число раз, которое встретила вершина x_i^* в списке, полученном по указанным выше путям.

Ещё один подход к определению перемычек основывается не на анализе графов, представляющих кластерные структуры в данных, а на построении приближающих многообразий. Приближение многомерных данных подходящими многообразиями малой размерности — бурно развивающаяся область современной методологии анализа, обработки и визуализации многомерных данных [5]. Предложенный ниже подход к определению перемычки целиком лежит в русле этого подхода. Интуитивно перемычка определяется как такое подмножество точек в (уже как-то кластеризованном множестве F_M), которое может быть эффективно приближено многообразием размерности заметно меньшей, чем то, которое хорошо приближает всё множество.

Перемычкой 1-го рода назовём множество S точек, удовлетворяющее условиям:

- 1) $S \subset F_M$,
- 2) S достаточно хорошо аппроксимируется одномерным многообразием, и
- 3) в S есть такая последовательность точек, что расстояние между соседними точками мало и начальная точка этой последовательности близка к одному кластеру, а конечная — к другому. При этом эти кластеры могут и совпадать.

В разности $F_M \setminus \hat{F}_M$ будем искать перемычку 1-го рода между кластерами ~ 1 и ~ 2 . В качестве базовой координаты $\sim r$ выберем ту, для которой разность $x_r^j - x_r^i$ для всех точек из $F_M \setminus \hat{F}_M$ наибольшая. Пусть это будет x_1 . Иными словами, это есть координата, проекция на которую потенциальной перемычки наибольшая.

В качестве центра потенциальной перемычки возьмём ту точку $a^0 \in F_M \setminus \hat{F}_M$, для которой x_1^0 наиболее близка к $1/2|a_1 - b_1|$, где a_1 и b_1 — x_1 -координаты центров кластеров ~ 1 и ~ 2 , соответственно. Центр нам нужен для того, чтобы запустить вычислительный процесс.

Зададим малый шаг δ вдоль оси x_1 . Рассмотрим множество

$$S_0 = \{P(x_1, x_2, \dots, x_n | x_1^0 \leq x_1 \leq x_1^0 + \delta)\} \subseteq F_M \setminus \hat{F}_M.$$

Это есть сечение «толщиной» δ множества $F_M \setminus \hat{F}_M$, ортогональное оси x_1 .

Сделаем кластеризацию этого сечения. Пусть V_0 — кластер в S_0 , содержащий или наиболее близкий к точке a_0 . Обозначим через d_0 диаметр этого кластера. Далее рассмотрим следующее сечение

$$S_1 = \{P(x_1, x_2, \dots, x_n | x_1^0 + \delta \leq x_1 \leq x_1^0 + 2\delta)\} \subseteq \mathbf{F}_M \setminus \hat{\mathbf{F}}_M.$$

Произведём кластеризацию сечения S_1 . Пусть V_1 — кластер в S_1 диаметра d_1 , удовлетворяющий следующим двум условиям: а) расстояние от его центра до центра кластера V_0 не превышает λ , где $\lambda \propto \delta$; б) расстояние от его центра до центра кластера 1 (или кластера 2) меньше, чем расстояние от центра кластера V_0 до центра кластера 1 (или кластера 2). Для определённости будем считать, что V_1 ближе к кластеру 2.

Аналогично построим сечения $S_2, S_3, \dots$ и кластеры $V_2, V_3, \dots$ диаметров $d_2, d_3, \dots$. Если на некотором шаге кластер V_1 стал достаточно близок к кластеру 2, то этот этап процесса прекращаем и начинаем двигаться в противоположную сторону, т.е. строим сечение

$$S_{-1} = \{P(x_1, x_2, \dots, x_n | x_1^0 - \delta \leq x_1 \leq x_1^0)\} \subseteq \mathbf{F}_M \setminus \hat{\mathbf{F}}_M$$

и так далее. Если же очередной кластер V_p , удовлетворяющий описанным выше условиям не существует, то и искомой перемиčky не существует. Если оба процесса достигли цели и $\max_k \{d_k\}$ достаточно мал, то объявляем полученное множество

$$V = \bigcup V_k$$

перемичкой 1-го рода.

Далее в множестве $\{\mathbf{F}_M \setminus \hat{\mathbf{F}}_M\} \setminus V$ можно искать следующую перемичку.

Можно привести примеры, когда предложенный метод не находит существующие перемиčky. Но это не удивительно. Не бывает же, например, универсального численного метода решения систем линейных уравнений.

Список литературы

1. Leskovec J., Rajaraman A., Ullman J. D. Mining of massive datasets. (2014) — Cambridge Univ. Press, 495 p.
2. Dongkuan Xu, Yingjie Tian (2015) A Comprehensive Survey of Clustering Algorithms. *Ann. Data. Sci.*, 2 (2): с.165 – 193.
3. Bondy J.A., Murty U. S. R. Graph theory with applications. (2007) — Elsevier. 270 p.
4. Girvan M., Newman M.E.J. (2002) Community structure in social and biological networks. *PNAS*, 99 (12): с.7821 – 7826.
5. Principal Manifolds for Data Visualization and Dimension Reduction (Eds. A. Gorban, B. Kégl, D.C. Wunsch, A. Zinovyev) (2008) Lect. Notes in Comp. Science and Engineering, v.58 — Springer-Verlag Berlin Heidelberg, XXIV, 340 p.

РАЗРАБОТКА МУЛЬТИАГЕНТНОЙ SMART-СИСТЕМЫ ДИСТАНЦИОННОГО ОБУЧЕНИЯ ЛЮДЕЙ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗРЕНИЯ⁴

Г.А. Самигулина¹, А.Т. Нюсупов², А.С. Шаяхметова³

^{1, 2, 3} Институт информационных и вычислительных технологий КН МОН РК,
лаборатория интеллектуальных систем управления и прогнозирования
galinasamigulina@mail.ru, moniumverse@otlook.com, asemshayakhmetova@mail.ru

Аннотация. Статья посвящена разработке мультиагентной smart-системы дистанционного обучения (ДО) для людей с ограниченными возможностями зрения (ЛОВЗ). Рассматриваются приложения на основе мультиагентных систем (МАС) в образовании, созданные с использованием платформы JADE (Java Agent Development Framework). Предложен комплексный подход ДО ЛОВЗ для получения качественного инженерного образования в лабораториях коллективного пользования (ЛКП) на современном оборудовании. Разработана архитектура мультиагентной smart-системы ДО ЛОВЗ на основе применения когнитивного, онтологического, статистического и интеллектуального подходов.

Введение. В настоящее время дистанционное обучение (ДО) является важной формой образования, которое эффективно развивается вместе с научно-техническим прогрессом [1]. Благодаря гибкости и универсальному подходу ДО стало конкурировать с традиционными системами обучения, предлагая новые методы модернизации образовательной деятельности на основе применения новейших технологий, последних достижений вычислительной техники и современных телекоммуникационных устройств [2,3]. Наибольший интерес представляют smart-системы дистанционного обучения и smart-технологии, которые обладают широкими возможностями для обеспечения высокого уровня получения знаний.

Актуально использование smart-систем ДО для людей с различными ограниченными возможностями, особенно для людей с патологиями зрения [4,5]. Эта категория людей нуждается в специализированных системах, так как ЛОВЗ сталкиваются с трудностями получения знаний в обычных образовательных учреждениях и традиционных системах ДО из-за особенностей восприятия информации. Важной задачей является организация специальной доступной среды ДО для людей с нарушениями зрения, так как данный вид обучения подразумевает длительную работу с компьютером, что сильно повышает нагрузку на зрительный аппарат и снижает уровень осознания учебной информации [6]. Возникает необходимость внедрения инновационных комплексных подходов, которые нацелены на повышение качества систем ДО ЛОВЗ. Наиболее удобно реализовывать данные задачи на основе мультиагентных систем (МАС). В таких системах решение получается автоматически в результате взаимодействия множества самостоятельных целенаправленных программных модулей – так называемых агентов. [7-10].

Для построения мультиагентной среды ДО используются различные агентные платформы (АП). Одной из широко распространенных программных сред для разработки МАС является JADE, написанная на языке Java [11]. Основой данной платформы является программная среда, без нее невозможно существование агентов.

⁴ Работа выполнена по гранту Комитета Науки Министерства Образования и Науки Республики Казахстан №ГР0215РК01472 (2015–2017гг.), по теме «Разработка информационной технологии, алгоритмов и программно-аппаратного обеспечения для интеллектуальных систем управления сложными объектами в условиях параметрической неопределенности».

Внутри среды формируются контейнеры, в которых функционируют агенты. После запуска каждый агент должен передать данные о себе в один из контейнеров, чтобы зарегистрироваться в системе. Далее программная среда будет наблюдать за работой всей системы и при необходимости производить требуемые управляющие воздействия на отдельных агентов. Между собой агенты обмениваются при помощи сообщений на языке ACL, который поддерживает современные стандарты передачи данных. Во время своей работы агент накапливает сообщения, которые приходят ему от других агентов и по очереди занимается их обработкой.

Разработано большое количество MAC на основе платформы JADE. В статье [12] описана MAC, которая предназначена для обеспечения безопасности мобильных устройств и защиты их от вредоносных программ. Особый упор делается на мониторинге SMS (Short Message Services) услуг. Агенты выполняют функцию поведенческого анализа путем выявления корреляции между подозрительными SMS-сообщениями и своими профилями. В работе [13] представлен агентно-ориентированный программный модуль для разработки и координации smart-объектов, которые представляют собой цифровые объекты информационной системы, дополненные вычислительными и накопительными возможностями. Smart-объектам присваиваются отдельные агенты, которые связываются между собой, могут выполнять общую организованную работу. В статье [14] показана мультиагентная модель приложения, которая применяется для индексации поиска и подачи рекомендательных учебных материалов, хранящихся на различных серверах. Данная модель разрабатывается с помощью агентной платформы JADE и стандарта метаданных LOM (Learning Object Metadata), которая имеет иерархическую структуру и обеспечивает быстрый доступ к заложенной информации. Технология мультиагентов повышает точность и релевантность поиска. В статье [15] показана система моделирования информационных процессов с использованием JADE. Используется библиотека, которая интегрируется в JADE и позволяет осуществлять правильное распределение ресурсов. В результате сравнительных тестов было доказано, что в практической реализации данное моделирование на основе агентов показывает наилучшие результаты.

Постановка задачи формулируется следующим образом: необходимо реализовать мультиагентную smart-систему на основе платформы JADE (Java Agent Development Framework) для дистанционного обучения людей с ограниченными возможностями зрения, с целью получения качественного инженерного образования в лабораториях коллективного пользования (ЛКП).

Введем следующее определение: в качестве лабораторий коллективного пользования рассматриваются любые лаборатории, имеющие полную инфраструктуру (современное оборудование, техническое и программное обеспечение) для осуществления дистанционного обучения и получения качественного инженерного образования людьми с ограниченными возможностями зрения.

Архитектура мультиагентной smart-системы дистанционного обучения людей с ограниченными возможностями зрения. При разработке мультиагентной smart-систем ДО ЛОВЗ созданы агенты на основе применения когнитивного, онтологического, статистического и интеллектуального подходов. В таблице приведены 5 контейнеров в которых помещены соответствующие агенты: контейнер 1 – менеджер агент (МА), контейнер 2 – персональный (ПА), когнитивный (КА) и интеллектуальный (ИА) агент, контейнер 3 – статистический (СА) и онтологический агент (ОА), контейнер 4 – агент помощи (АП) и здоровья (АЗ), контейнер 5 – агент ЛКП (АЛ) и государственный агент (ГА).

Таблица контейнеров и агентов smart-системы ДО ЛОВЗ

Контейнер	Агент	Функции агента
1	2	3
контейнер 1	1. менеджер агент (МА)	– отслеживание потоков сообщений от других агентов; – организация режима подачи информации.
контейнер 2	1. персональный агент (ПА)	– регистрация персональных данных и сбор индивидуальных характеристик (признаков) ЛОВЗ; – организация траектории обучения ЛОВЗ с учетом дефектов зрения; – подбор нужных учебных материалов.
	2. когнитивный агент (КА)	– выявление уровня восприятия информации ЛОВЗ; – реализация когнитивных методик обучения.
	3. интеллектуальный агент (ИА)	– определение уровня обучения на основе нечеткой логики; – прогнозирование результатов обучения ЛОВЗ с помощью нейро-нечеткой сети.
контейнер 3	1. статистический агент (СА)	– учет динамики ДО пользователей; – запись логов событий и системных ошибок; – выделение информативных признаков ЛОВЗ на основе факторного анализа.
	2. онтологический агент (ОА)	– построение OWL модели ЛОВЗ; – структурирование входных и выходных данных smart-системы.
контейнер 4	1. агент помощник (АП)	– помощь студентам в выборе материалов обучения посредством подсказок и уведомления.
	2. агент здоровья (АЗ)	– предоставление индивидуального режима работы на компьютере с учетом особенностей зрения; – подбор комплекса упражнения для глаз с учетом заболевания зрения.

Продолжение таблицы

контейнер 5	1. агент ЛКП (АЛ)	– подключение ЛОВЗ к ЛКП; – доступ к современному оборудованию.
	2. государственный агент (ГА)	– соединение smart-системы ДО ЛОВЗ с государственными учреждениями; – предоставление юридической информации пользователям; – доступ к нормативно-правовой базе.

Для программирования конфигурации созданных агентов используется среда, включающая в себя классы и библиотеки платформы JADE. Агенты функционируют в вычислительной среде, которая постоянно изменяется под воздействием деятельности ЛОВЗ в системе.

Заключение. Разработана мультиагентная smart-система ДО ЛОВЗ для дистанционного обучения ЛОВЗ, позволяющая получить качественное инженерное дистанционное образование на основе построения индивидуальной траектории обучения ЛОВЗ с учетом интеллектуальных и психофизических особенностей восприятия и осознания информации. Разработанная smart-система ДО ЛОВЗ имеет следующие достоинства:

- выявление физиологических интеллектуальных и психофизиологических особенностей восприятия и осознания информации ЛОВЗ;
- построение различных OWL моделей ЛОВЗ и систематизация предметной области, обработка статистических данных и выделение информативных признаков ЛОВЗ;
- интеллектуальный анализ данных и прогнозирование результатов обучения ЛОВЗ, формирование специальной среды восприятия ЛОВЗ;
- учет подачи и расположения учебной информации на экране монитора в зависимости от дефектов зрения, возможность выбора цветовых схем, наилучших для людей с различными заболеваниями;
- доступ к современному оборудованию и техническим средствам для ЛОВЗ, с целью изучения дисциплин по инженерным специальностям;

Таким образом, развитие и применение мультиагентных технологий для дистанционного обучения с использованием современных методов искусственного интеллекта позволяет разрабатывать эффективные интеллектуальные информационные системы, востребованные людьми с ограниченными возможностями зрения.

Список литературы

1. Olaf Zawacki-Richter. Distance Education // Encyclopedia of Educational Philosophy and Theory. — Singapore: Springer Science. — 2015. — P.1 – 7.
2. Zhang HongYan. Study on Learning Performance Evaluation of Distance Continuing Education / Zhang HongYan Wang JiKui // Advanced Technology in Teaching. — Berlin: Springer. — 2013. — P. 255-260.

3. Sônia Regina. Interaction in Distance Education: Student, Teaching Material, Information Technology and Communication / Sônia Regina, Gouvêa Rezende, Valter Gomes Campos, Pollyana dos Reis Pereira Fantone, Melca Moura Brasil // HCI International. Posters' Extended Abstracts. — Berlin: Springer – Heidelberg. — 2013. — P. 76-79.
4. Samigulina G. A. The information system of distance learning for people with impaired vision on the basis of artificial intelligence approaches / G. A. Samigulina, A. S. Shayakhmetova // Proceedings of the II International conference on Smart Education and E-learning. — Italy, Sorrento: Springer. — 2015. — P. 255-265.
5. Rabia Jafri. Tablet-Based Application for the Visually Impaired for Detecting and Recognizing Objects in a Home Environment / Rabia Jafri, Syed Abid Ali A Multimodal // Computers Helping People with Special Needs. — Switzerland: Springer International Publishing. — 2014. — P. 356-359.
6. Samigulina G.A. Smart-system of distance learning of visually impaired people based on approaches of artificial intelligence / G.A. Samigulina, A.S. Shayakhmetova // J. Open Engineering. — 2016. — № 6. — 359-366.
7. Max Gath. Autonomous, Adaptive, and Self-Organized Multiagent Systems for the Optimization of Decentralized Industrial Processes / Max Gath, Otthein Herzog, Stefan Edelkamp // Intelligent Agents in Data-intensive Computing. — Switzerland: Springer International Publishing. — 2016. — P. 71-98.
8. Самигулина Г.А. Обзор интеллектуальных систем для дистанционного образования, построенных на основе мультиагентного подхода / Г.А. Самигулина, А.Т. Нюсупов // Проблемы информатики. — Новосибирск. — 2017. — № 2. — С. 24-37.
9. Stamatis P. APLe: Agents for Personalized Learning in Distance Learning / P. Stamatis, I. Panagiotopoulos, C. Goumopoulos, A. Kameas // Computer Supported Education. — Switzerland: Springer International Publishing. — 2016. — P. 37-56.
10. Samigulina G. A. Multi-agent Smart-system of distance learning for people with vision disabilities / G. A. Samigulina, A.T. Nyussupov, A.S. Shayakhmetova // Proceedings of the IV International conference on Smart Education and E-learning. — Portugal, Algarve: Springer. — 2017. — P. 154-166.
11. JADE Architecture Overview. <http://www.jade.tilab.com/>
12. Alzahrani A.J. A Multi-agent System for Smartphone Intrusion Detection Framework / A.J. Alzahrani, A.A. Ghorbani // Proceedings of the 18th Asia Pacific Symposium on Intelligent and Evolutionary Systems. — Vol. 1. — Switzerland: Springer International Publishing. — 2015. — P. 101-113.
13. Giancarlo Fortino. An Agent-Based Middleware for Cooperating Smart Objects / Giancarlo Fortino, Antonio Guerrieri, Michelangelo Lacopo, Matteo Lucia, Wilma Russo // Highlights on Practical Applications of Agents and Multi-Agent Systems. — Berlin: Springer, Heidelberg. — 2013. — P. 387-398.
14. Ronaldo Lima RochaCampos. Multiagent Based Recommendation System Model for Indexing and Retrieving Learning Objects / Ronaldo Lima RochaCampos, Rafaela Lunardi Comarella, Ricardo Azambuja Silveira // Highlights on Practical Applications of Agents and Multi-Agent Systems. — Berlin: Springer, Verlag. — 2012. — P. 328-339.
15. Henrique Lopes Cardoso. SAJaS: Enabling JADE-Based Simulations // Transactions on Computational Collective Intelligence XX. — Berlin: Springer, Heidelberg. — 2013. — P. 387-398.

ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ СИНТЕЗА НОВЫХ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ СУЛЬФАНИЛАМИДОВ⁵

З.И. Самигулина¹, Т.И. Самигулин²

¹Институт информационных и вычислительных технологий КН МОН РК,
Лаб. «Интеллектуальные системы управления и прогнозирования»,
zarinasamigulina@mail.ru

²Казахский национальный исследовательский технический университет
им. К.И. Сатпаева, ***timur.samigulin@yandex.kz***

В настоящее время процесс создания новых лекарственных препаратов является трудоемкой задачей, требующей больших временных и финансовых затрат. Последние достижения в области фармакологии, биоинформатики и хемоинформатики, а так же возможности компьютерного моделирования позволяют разрабатывать лекарственные препараты нового поколения. Междисциплинарная направленность исследований и огромное количество открытых структур химической информации создают дополнительные сложности при проведении исследований [1].

Широко известен метод QSAR (Quantitative Structure-Activity Relationship), который позволяет производить исследования зависимости между структурами химических соединений и их биологической активностью. Для поиска количественных соотношений «структура/свойство» хорошо зарекомендовали себя современные методы искусственного интеллекта и статистического анализа данных, позволяющие предсказывать физические, химические и др. свойства химических соединений. К ним относятся: искусственные нейронные сети [2], генетические алгоритмы [3], искусственные иммунные системы [4,5], метод главных компонент (Principle Component Analysis, PCA), регрессионный анализ (Partial least squares regression, PLS), метод опорных векторов (SVM) [6], SOM – сети, модель Random Forest (RF) [7], метод ближайшего соседа (k-NN) [8] и т.д. Поскольку в процессе исследования зависимости «структура/свойство» химических соединений необходимо обрабатывать огромный объем данных разработаны современные специализированные программные продукты для автоматизированного ведения научных исследований, такие как: WEKA, Rapid Miner, LIBSIM, Bioclipse, RStudio и т.д. Таким образом, применение специализированного программного обеспечения при синтезе новых лекарственных препаратов позволяет осуществлять высокопроизводительное компьютерное моделирование на основе современных методов искусственного интеллекта и визуализацию данных.

Постановка задачи формулируется следующим образом: необходимо осуществить исследование зависимости «структура/свойство» химических соединений сульфаниламидной группы на базе современных интеллектуальных алгоритмов. Провести сравнительный анализ полученных результатов.

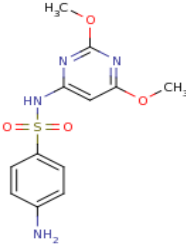
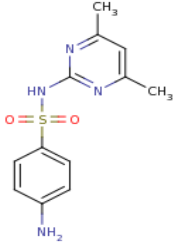
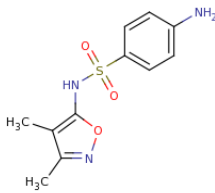
Рассмотрим класс противомикробных сульфаниламидных лекарственных препаратов. В настоящее время открыто и изучено более 6000 производных данной группы, однако практическое применение находят около 20 препаратов. Разработана база данных дескрипторов сульфаниламидов на основе современных банков данных химической информации: Mol-Instincts содержит более 2.85 млн. компонентов, PubChem около 93 553 257 компонентов по состоянию на 2017 г. Поиск по базе данных

⁵ Работа выполнена по гранту КН МОН РК № ГР 0115РК00549 (2015-2017 гг.) по теме проекта: "Компьютерный молекулярный дизайн лекарственных препаратов на основе иммуносетевого моделирования".

осуществлялся с помощью спецификации SMILES (Simplified Molecular Input Line Entry Specification). В таблице 1 представлен фрагмент базы данных сульфаниламидов, их химическая формула и 2D структура.

Таблица 1

Фрагмент базы данных сульфаниламидов

Химическая формула	Спецификация SMILES	2D структура
2	3	4
$C_{12}H_{14}N_4O_4S$	<chem>COc2cc(NS(=O)(=O)c1ccc(N)cc1)nc(OC)n2</chem>	
$C_{12}H_{14}N_4O_2S$	<chem>Cc2cc(C)nc(NS(=O)(=O)c1ccc(N)cc1)n2</chem>	
$C_{11}H_{13}N_3O_3S$	<chem>Cc2noc(NS(=O)(=O)c1ccc(N)cc1)c2C</chem>	
...	...	...

Сульфаниламиды описываются с помощью дескрипторов различного уровня:

1. Структурные дескрипторы (число атомов; относительное количество атомов C, H, O, N, S; число одинарных связей; число ароматических связей; молекулярный вес; Гравитационный; средний молекулярный вес и т.д.). Общее число характеристик - 58 дескрипторов.

2. Топологические дескрипторы (индекс молекулярной подвижности Кира; среднее значение содержания информации; индекс Балабана; индекс Загреба; квадратичный индекс; простой топологический индекс Наруми; молекулярный топологический индекс Шульца; молекулярный топологический индекс Гутмана; индекс симметрии Кира и т.д.). Общее число характеристик - 114 дескрипторов.

4. Расстояние и расчет траектории (молекулярное расстояние; общее расстояние и т.д.). Общее число характеристик - 47 дескрипторов.

3. Индекс связывания (индекс валентного связывания; модифицированный индекс связывания Рандика и т.д.). Общее число характеристик – 33.

4. Информационный индекс (информация о размере молекулы; общий индекс информации атомного состава; среднее содержание информации о расстоянии; индекс сложности вершин графа; информация о количестве связей; содержание структурной информации и т.д.). Общее число характеристик - 47 дескрипторов.

5. Топологические индексы заряда. Общее число характеристик - 21 дескриптор.
6. Геометрические дескрипторы (молекулярное значение; молекулярная поверхность; 3D индекс Винера; 3D индекс Балабана; 3D Харари индекс; среднее геометрическое расстояние; индекс дистанции и т.д.). Общее число характеристик - 56 дескрипторов.
7. RDF - дескрипторы. Общее число характеристик - 150 дескрипторов.
8. 3D-MoRSE дескрипторы. Общее число характеристик - 160 дескрипторов.
9. 2D двоичные отпечатки. Общее число характеристик - 100 дескрипторов.
10. Квантово-химические характеристики (число уровней электронов; min. индекс нуклеофильной реактивности и т.д.). Общее число характеристик - 155 дескрипторов.
11. Электростатические дескрипторы (max. частичный заряд; параметр полярности; топографический электронный индекс и т.д.). Общее число характеристик - 65 дескрипторов и т.д. [9].

В таблице 2 представлен фрагмент базы данных дескрипторов сульфаниламидов, содержащей в общей сложности 32080 экземпляра данных.

Таблица 2

Фрагмент базы данных дескрипторов сульфаниламидов

Дескриптор	Sulfadiazine	Sulfadimidine	Sulfafurazole	...	Sulfaperin
Number of atoms	27.0000	33.0000	31.0000	...	30.0000
Relative number of C atoms	0.370400	0.363600	0.354800	...	0.366700
Relative number of H atoms	0.370400	0.424200	0.419400	...	0.400000
Number of single bonds	16.0000	24.0000	26.0000	...	22.0000
Relative number of single bonds	0.571400	0.705900	0.812500	...	0.709700
...	...	...	...	...	...
Topographic electronic index	12.0000	10.0000	6.0000	...	9.0000

Исследования базы данных (БД) дескрипторов сульфаниламидов осуществлялись на базе программных продуктов WEKA и Rapid Miner. Визуализация БД сульфаниламидов представлена на рисунке 1.

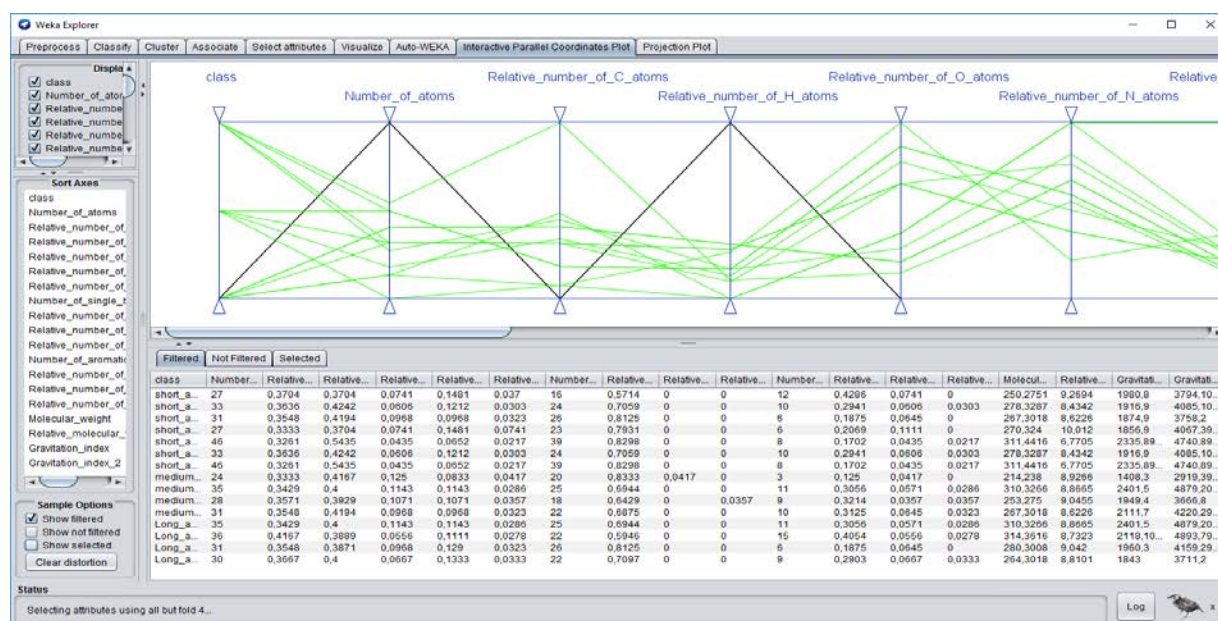


Рис. 1. Визуализация БД дескрипторов сульфаниламидов в среде WEKA

Рассмотрим алгоритм интеллектуального анализа БД дескрипторов сульфаниламидов для прогнозирования зависимости «структура/свойство» лекарственных препаратов.

Алгоритм.

1. Загрузка БД дескрипторов химического вещества в среду для автоматизированного ведения научных исследований (WEKA или Rapid Miner).

2. Предварительная обработка БД дескрипторов для снижения размерности исследуемых данных на основе алгоритма Random Forest [10].

3. Формирование БД, состоящей из информативных дескрипторов по результатам предварительной обработки данных.

4. Решение задачи классификации.

4.1 Формирование классов экспертами.

1 класс – длительность вещества менее 10 часов, сульфаниламиды малой продолжительности действия (Sulfaisodimidine, Sulfamethizole, Sulfadimidine, Sulfapyridine, Sulfafurazole и т.д.).

2 класс – длительность вещества от 10 до 24 часов, сульфаниламиды средней продолжительности действия (Sulfamethoxazole, Sulfadiazine, Sulfamoxole и т.д.).

3 класс – длительность вещества от 24 до 48 часов, сульфаниламиды длительной продолжительности действия.

4.2 Решение задачи классификации на основе метода «ближайшего соседа».

4.2.1 Оценка эффективности алгоритма.

4.3 Решение задачи классификации с помощью алгоритма метода опорных векторов.

4.3.1 Оценка эффективности алгоритма.

4.4 Сравнительный анализ полученных результатов. Принятие решений.

Реализация представленного алгоритма осуществлялась с помощью создания графической модели [11] в среде Rapid Miner (рисунк 2).

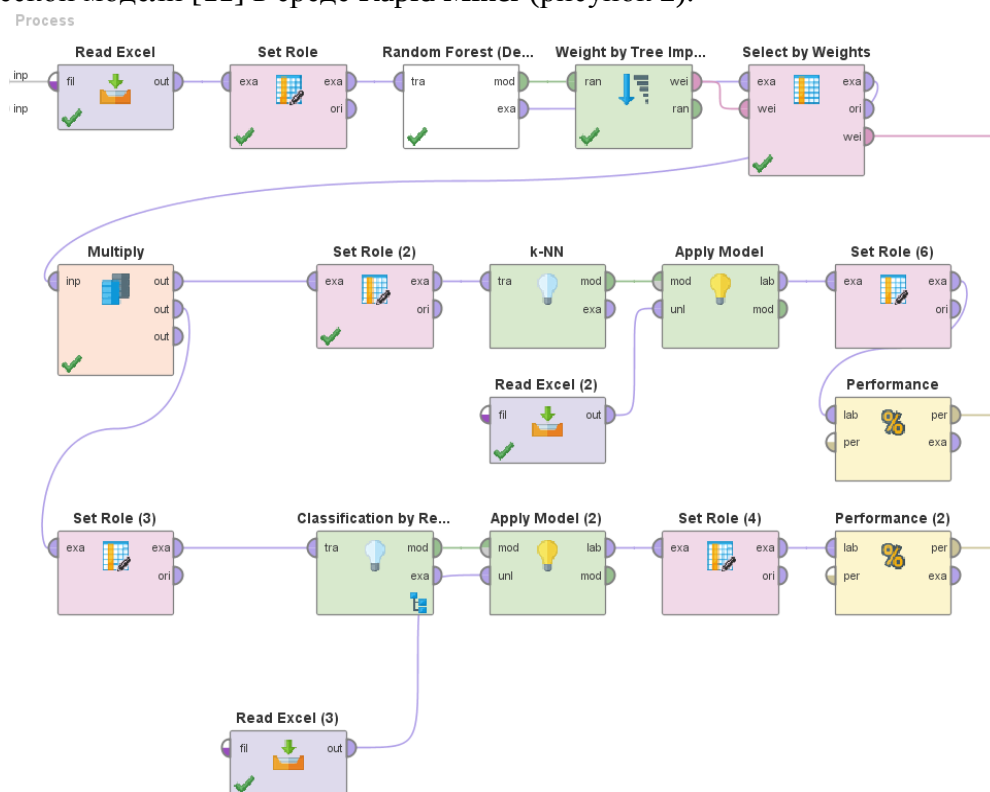


Рис. 2. Графическая модель интеллектуального анализа дескрипторов сульфаниламидов

Для ранжирования переменных по степени значимости и формирования БД информативных дескрипторов сульфаниламидов используется оператор «Weight by tree importance», чем больше параметр «weight», тем значимее дескриптор для описания свойств химических соединений. В таблице 3 представлен фрагмент БД информативных дескрипторов сульфаниламидов.

Таблица 3

Фрагмент БД дескрипторов сульфаниламидов после обработки на основе алгоритма RF

№	D1_RF	D2_RF	D3_RF	...	D74_RF
1	12.01	20.50	37.80	...	664
2	12.69	20.49	37.81	...	665
3	22.10	20.50	37.81	...	669
4	18.75	20.36	37.18		662.11
5	19.65	20.34	37.83		662.32
...	...	...	...	...	...
16	82.50	20.51	37.40	...	713

По результатам обработки БД дескрипторов сульфаниламидов снизилась до 1184 экземпляров данных. Решение задачи классификации методами k-NN и SVM осуществлялось на основе уменьшенной БД. В таблице 4 представлены результаты оценки эффективности работы алгоритмов с помощью оператора «Perfomance».

Таблица 4

Результаты интеллектуального анализа данных в среде Rapid Miner

Алгоритм исследования	Random Forest	k-NN	SVM
Время моделирования	0.03	0.02	0.01
Оценка эффективности работы алгоритма	94%	88%	91%

Сравнительный анализ результатов моделирования показывает, что метод «ближайшего соседа» обладает следующими достоинствами: простотой использования результатов моделирования, полученные решения не уникальны, а могут применяться в других случаях. В качестве недостатков можно отметить: вычислительную трудоемкость за счет необходимости полного перебора обучающей выборки данных и возможность применения для обработки БД небольшой размерности. Исследуемый метод опорных векторов подходит для работы с реальными данными, а так же использует линейный классификатор для работы с нелинейно-разделяемыми данными. Недостатком является использование для классификации не всего множества образцов, а лишь их небольшой части, которая находится на границах областей [12].

Таким образом, применение современных методов искусственного интеллекта для решения задачи предварительной обработки данных и классификации с помощью специализированных программных продуктов для автоматизированного ведения научных исследований (WEKA и Rapid Miner) является перспективным направлением в области прогнозирования зависимости «структура/свойство» химических соединений с целью создания новых лекарственных препаратов.

Список литературы

1. Головки Ю.С., Ивашкевич О.А., Головки А.С. Современные методы поиска новых лекарственных средств // Вестник БГУ. – 2012. - №1. – С. 7-15.
2. Priest A.C., Williamson A.J., Cartwright H.M. The Applications of Artificial Neural Networks in the Identification of Quantitative Structure-Activity Relationships for Chemotherapeutic Drug Carcinogenicity //Advances in Intelligent Data Analysis IX. – 2010. - №6065. – P. 137-146.
3. Sukumar N., Prabhu G., Saha P. Applications of Genetic Algorithms in QSAR/QSPR Modeling //Applications of Metaheuristics in Process Engineering. – 2014. – P. 315-324.
4. Ivanciuc O. Artificial Immune System Classification of Drug-induced Torsade de Pointes with AIRS (Artificial Immune Recognition System) // Journal of Molecular Design. – 2006. - №5. – P. 488-502.
5. Samigulina G.A., Samigulina Z.I. Immune Network Technology on the Basis of Random Forest Algorithm for Computer-Aided Drug Design // 5th International Work-Conference, IWBBIO 2017, Bioinformatics and Biomedical Engineering. - Granada, Spain, 2017. – Vol. 10208. – P. 50-61.
6. Liu F., Zhou Z. A new data classification method based on chaotic particle swarm optimization and least square-support vector machine // Chemometrics and intelligent laboratory systems. – 2015. – P.147-156.
7. Ko G.M., Reddy S., Kumar S., Bailey B.A., Garg R. A Random Forest Model for the Analysis of Chemical Descriptors for the Elucidation of HIV-1 Protease Protein-Ligand Interactions // Applied Computational Science and Engineering Student and Computational Science Curriculum Development (ACSESS), SDSU. – 2010. – P. 1-6.
8. Chikh M.A., Saidi M. Diagnosis of Diabetes Diseases Using an Artificial Immune Recognition System2 (AIRS2) with Fuzzy K-nearest Neighbor // Journal of Medical Systems. – 2012. – Vol. 36, №5. – P. 2721-2729.
9. Sharma B.K., Pilana P., Sarbhai K., Singh P., Prabhakar S.Y. Chemometric descriptors in modeling the carbonic anhydrase inhibition activity of sulfonamide and sulfamate derivatives // Molecular Diversity. – 2010. – №2. – P. 371-384.
10. Самигулина Г.А., Самигулина З.И. Построение оптимальной иммуносетевой модели для компьютерного молекулярного дизайна сульфаниламидов на основе алгоритма Random Forest //Вестник ХПИ, 2016. - №21(1193). - С. 102-109.
11. Самигулина Г.А., Самигулина З.И. Разработка графических моделей для интеллектуального анализа дескрипторов сульфаниламидов при синтезе новых лекарственных препаратов //Материалы научной конференции института информационных и вычислительных технологий «Современные проблемы информатики и вычислительных технологий». – 2017. – С. 223 – 228.
12. Чубукова И.А. Data Mining. – М.: Бином, 2008. – 382 с.

СТРУКТУРА ГЕНОМОВ ЦИАНОБАКТЕРИЙ

М.Ю. Сенашова¹, М.Г. Садовский^{1,2}

¹Институт вычислительного моделирования СО РАН,
660036, Красноярск, Академгородок
msen@icm.krasn.ru, msad@icm.krasn.ru

²Сибирский Федеральный университет,
Институт фундаментальной биологии и биотехнологии,
660041, Красноярск, пр.Свободный, 79

Введение

Изучение особенностей и деталей структуры нуклеотидных последовательностей является важнейшей задачей современной биологии. Такого рода исследования ведутся в двух аспектах — структурно-функциональном и эволюционном. Выявление связи структуры и функции представляет собой классическую проблему молекулярной и системной биологии и несмотря на обширный поток публикаций и работ в данном направлении, она всё ещё далека от завершения. Более того, исследователи выявляют всё новые и новые структурные элементы либо новые виды и формы взаимодействий и взаимоотношений между структурными элементами биологических макромолекул, а развитие техники и инструментов исследований лишь усугубляет эту ситуацию.

Эволюционный аспект таких исследований вполне понятен: изучая особенности структуры биологических макромолекул (ДНК либо белковых последовательностей) у разных организмов, можно точнее составить картину того, как эволюционировали те или иные биологические системы — от конкретных видов (рас, штаммов и т. п.) до экосистем и глобальных сообществ.

Такого рода исследования всегда затруднены выбором и качеством того биологического материала, который берётся в рассмотрение. Дело даже не в ошибках (к примеру) секвенирования и/или аннотирования генетических последовательностей, неизбежных во многих случаях, а в большой сложности таких объектов, как геномы либо отдельные хромосомы. Фактически приходится анализировать целую триаду характеристик: структуру, функцию и филогению. Эти элементы триады плотно переплетены друг с другом и оказывают друг на друга очень сильное влияние, которое далеко не всегда удастся выделить в качестве отдельного и независимого фактора.

Прокариотические организмы с этой точки зрения являются более удобными объектами для исследования, чем эукариотические; геном бактерий заметно короче генома эукариот и всегда представлен одной хромосомой. Своего рода расплатой за такое удобство является заметная трудность в определении филогении бактерий, особенно для таксонов высокого уровня.

Исследования структур в генетических последовательностях также является важной задачей, осложняющейся чрезвычайно большим разнообразием структур, которые можно найти и выделить в молекулах ДНК, даже если не обращать внимания на их химические свойства. Настоящая работа посвящена изучению структур в геномах цианобактерий

Под структурой мы будем понимать различие (либо подобие) статистических свойств отдельных формально выделяемых фрагментов генома на уровне триплетов. Иными словами, структура в рамках настоящей работы — это взаимное расположение различных (формально выделяемых) фрагментов генома сравнительно небольшой длины в пространстве частот триплетов, которые подсчитываются в пределах указанных фрагментов; подробности представлены в разделе 2.

Данный подход к изучению структурированности геномов и их связи с их GC-контентом был впервые предложен в работах Горбаня с соавторами [1-3]. Этот же подход

(с небольшой модификацией) используется и нами; один из мотивов использования метода, предложенного в указанных работах — теория симбиогенеза [4, 5]: поскольку согласно этой теории современные хлоропласты и цианобактерии имеют общего предка. Если эта теория верна, то можно надеяться найти какие-то признаки подобия структур, выделяемых в бактериальных геномах и в геномах хлоропластов. Анонсируя основной результат, скажем сразу, что были обнаружены существенные различия, а не подобие. Из этого не следует, что теория симбиотического происхождения хлоропластов неверна; из этого следует, что в ходе эволюции этих двух генетических изолированных систем произошла сильная дивергенция.

Материалы и методы

Введём основные понятия. Мы будем рассматривать генетическую последовательность длины L , состоящую из символов алфавита $\mathfrak{M} = \{A, C, G, T\}$. Если последовательность содержит символы, отличающиеся от символов алфавита $\mathfrak{M}$, то такие символы из последовательности удаляются, а длина последовательности уменьшается на число таких символов. Для этой последовательности мы будем составлять частотный словарь толщины 3. Частотный словарь толщины 3 символьной последовательности, соответствующей ДНК — это список всех троек $v_1 v_2 v_3$ идущих подряд нуклеотидов с указанием частот этих троек; всего может быть 64 триплета. Частота f_ω — это отношение числа копий n_ω данного слова к общему числу всех триплетов N , где N — сумма всех n_ω :

$$f_\omega = \frac{n_\omega}{N} \quad (1)$$

Всякий частотный словарь W_3 отображает геном в 64-мерное метрическое пространство; близость двух геномов задается естественным образом — например, как близость двух точек в Евклидовой метрике:

$$\rho(W_3^1, W_3^2) = \sqrt{\sum_{\omega=AAA}^{TTT} (f_\omega^1 - f_\omega^2)^2}. \quad (2)$$

Один из 64 триплетов исключался, поскольку сумма всех частот в словаре равна 1, что порождает линейную связь, которая будет давать ложный сигнал при статистической обработке (корреляционном анализе, определении главных компонент и т. п.). Формально исключить можно любой триплет, однако есть несколько эвристических подходов для исключения. Первый подход состоит в том, чтобы исключить самый большой по значению частоты триплет, особенно если значение его частоты на порядок (или около того) превосходит значение частоты следующего за ним (по этой величине) триплета.

Второй подход состоит в том, чтобы исключать тот триплет, для которого стандартное отклонение, наблюдаемое по анализируемому набору частотных словарей, является минимальным: такой триплет даёт наименьший вклад в различимость объектов (в предельном случае, когда стандартное отклонение равно 0, различий по этому триплету вовсе нет). Мы использовали второй подход. Таким образом, рассматриваемое нами пространство точек становится 63-х мерным. В наших исследованиях, в основном, исключались триплеты GCG и CGC, подробности см. в разделе 3.

Для выявления структуры в генетической последовательности проводилась предварительная обработка, которая ставила в соответствие данной последовательности множество точек в 64-мерном пространстве триплетов. Делалось это следующим образом: последовательность сканировалась окном длины Δ с шагом t . Для каждого

положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$, соответствующий i -ой точке в 64-мерном пространстве. Кроме того, с каждой точкой в 64-мерном пространстве связывались следующие параметры: номер центрального символа рассматриваемого участка и относительная фаза.

Номер центрального символа участка совпадает с номером этого символа в последовательности. Относительная фаза определяется в зависимости от того, попал рассматриваемый участок в кодирующую или не кодирующую область последовательности. Участок относится к кодирующим, если он целиком попадал в кодирующую область последовательности. Если участок относится к не кодирующим, то соответствующая ему точка помечается символом J . Для кодирующего участка возможны 6 вариантов маркировки: $B_0, B_1, B_2, F_0, F_1, F_2$. Если кодирующий участок в последовательности аннотирован как считывающийся в прямом направлении, то для него вычислялся остаток от деления на 3 разности номеров центрального символа участка и первого символа кодирующей области, к которой он относится. В соответствии с величиной остатка от деления точка помечалась символом B_0, B_1 или B_2 . Если участок аннотирован как считывающийся в обратном направлении, то вычислялся остаток от деления на 3 разности номеров последнего символа кодирующей области, к которой относится участок, и центрального символа участка. В зависимости от значения остатка от деления точка помечалась символами F_0, F_1 или F_2 . Для всех генетических последовательностей длина рамки считывания $\Delta = 603$, шаг $t = 11$. Такой выбор рамки считывания обусловлен следующими соображениями: во-первых, такая длина фрагмента близка к характерной длине гена; во-вторых, уменьшение рамки считывания хотя бы вдвое приводит к значительному вырождению частотных словарей. Это означает, что значительное (до трети) число всех слов в частотном словаре, соответствующем выделенному фрагменту, имеют нулевые частоты либо представлены в одной копии; тем самым, различимость фрагментов по этим триплетам отсутствует. В-третьих, увеличение длины выделяемого фрагмента до $\Delta \approx 1500$ не меняет качественно картину, в то время как увеличение Δ до 3000 и выше нуклеотидов приводит к образованию большого числа четко отделенных друг от друга плотных кластеров, в которых представлены фрагменты, относящиеся к разным относительным фазам. Косвенно это свидетельствует о том, что фрагменты длиной 3000 нуклеотидов и выше почти всегда содержат две кодирующие области.

По полученному множеству точек в программе *VidaExpert* (<http://bioinfo-out.curie.fr/projects/vidaexpert/>) строился вид данных в пространстве первых трёх главных компонент, вычисленных для данного 63-мерного пространства. Рассматривались проекции на плоскость пространства главных компонент. Точки данных были раскрашены в соответствии с их относительной фазой. Точки, соответствующие не кодирующим участкам, изображены на рисунках коричневым цветом, точки, соответствующие относительным фазам B_0 и F_0 , изображены темно малиновым и светло малиновым цветом, точки, соответствующие относительным фазам B_1 и F_1 , изображены темно зеленым и светло зеленым цветом, а точки, соответствующие относительным фазам B_2 и F_2 , изображены темно желтым и светло желтым цветом.

Результаты

Гипотеза о происхождении хлоропластов от одноклеточных свободноживущих фотосинтезирующих бактерий позволяет ожидать, что структура данных цианобактерий будет подобна аналогичной структуре хлоропластов. Мы рассмотрели структуры для геномов цианобактерий, депонированных в EMBL-банке, а именно *Microcystis aeruginosa* NIES-843, *Nostoc* sp. PCC 7107, *Pleurocapsa* sp. PCC 7327, *Chroococcidiopsis thermalis* PCC 7203, *Gloeocapsa* sp. PCC 7428, *Anabaena cylindrica* PCC 7122, *Synechocystis* sp. PCC 6714. Эти цианобактерии относятся к трем порядкам: Chroococcales, Nostocales, Pleurocapsales. На Рис. 1 показан характерный вид структуры генома цианобактерий в пространстве первых трех главных компонент. Как видно из рисунка, структура данных цианобактерий представляет собой своеобразные клубки из цепочек, точки в которых расположены в той же последовательности, в которой соответствующие фрагменты расположены в геноме.

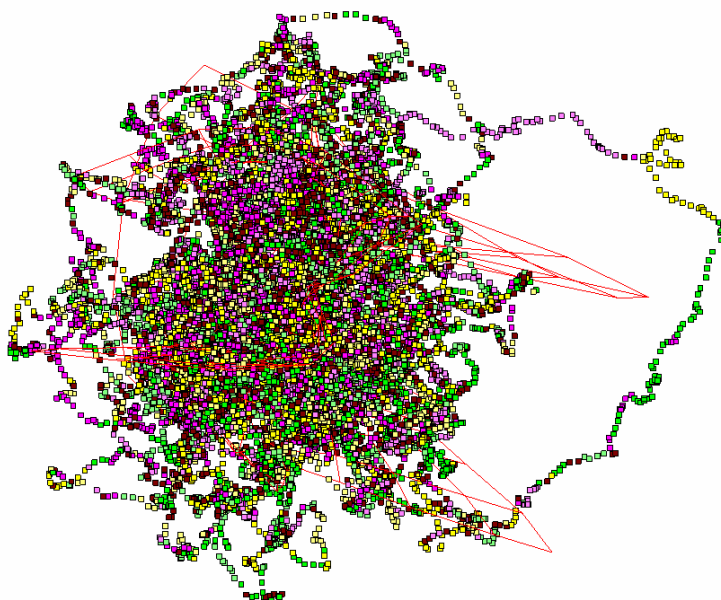


Рис.1. Структура данных *Microcystis aeruginosa* в пространстве главных компонент.

Кроме того, было обнаружено, что «клубок», соответствующий структуре цианобактерий в пространстве главных компонент состоит из одной или из трех нитей, в зависимости от величины шага t . В случае кратности шага трем, клубок состоит из одной нити, в противном случае нитей три. На рис.2 показаны фрагменты геномов *Microcystis aeruginosa* и *Nostoc*, включающих первую тысячу точек в пространстве первых трех главных компонент для $\Delta = 60003$ и $t = 303$. На рисунке видно, что структура генома представляет собой одну нить. На рис. 3 показаны фрагменты этих же геномов для $\Delta = 60003$ и $t = 91$.

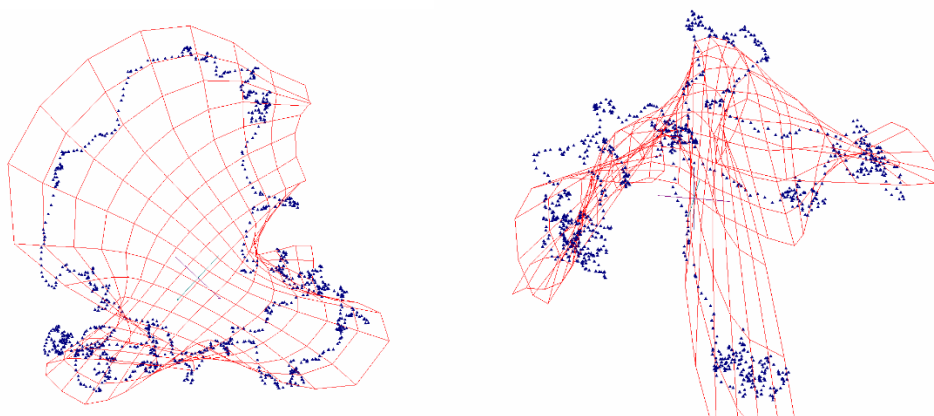


Рис.2. Структура геномов *Microcystis aeruginosa* и *Nostoc* при шаге t , кратном трем.

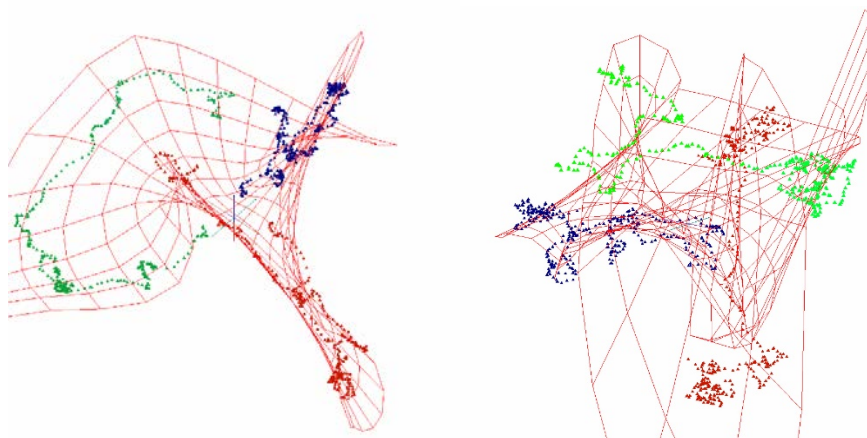


Рис.2. Структура геномов *Microcystis aeruginosa* и *Nostoc* при шаге t , не кратном трем.

Структура данных цианобактерий существенно отличается от структуры данных хлоропластов. Кроме того, структура данных цианобактерий существенно отличается и от структуры иных бактерий. Семикластерная структура фрагментов генома у цианобактерий отсутствует. Это также выделяет цианобактерии среди других бактерий и подчеркивает их особенности, первой среди которых выделяется способность к фотосинтезу. В частности, это наблюдение может свидетельствовать об очень древнем расхождении современных цианобактерий и хлоропластов от общего предка и о весьма сложных путях их эволюции.

Список литературы

1. Gorban A.N., Zinovyev A. Yu., Popova T.G., 2003. Seven clusters in genomic triplet distributions. // arXiv:cond-mat/0305681 [pdf]
2. Gorban A.N., Zinovyev A. Yu., Popova T.G., 2003. Seven clusters in genomic triplet distributions // Silico Biology. V.3. № 4. P. 471-482.
3. Gorban A.N., Zinovyev A. Yu., Popova T. G., 2005. Four basic symmetry types in the universal 7-cluster structure of microbial genomic sequences // Silico Biology. V.5. № 3. P.265-282.
4. Мережковский К.С. 1909. Теория двух плазм как основа симбиогенезиса, нового учения о происхождении организмов. Казань: Изд-во Импер. ун-та, 102 с.
5. Mereschkovsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche // Biol. Zentr.-Bl. Bd. 85. № 18. S. 593–604.

АНАЛИЗ КЛАСТЕРНОЙ СТРУКТУРЫ ГЕНОМОВ ХЛОРОПЛАСТОВ ВЫЯВЛЯЕТ ИХ ОТЛИЧИЯ ОТ ГЕНОМОВ ПРОКАРИОТ

М.Ю. Сенашова¹, М.Г. Садовский^{1,2}

¹Институт вычислительного моделирования Сибирского отделения Российской академии наук, ²Сибирский федеральный университет, институт фундаментальной биологии и биотехнологии, msen@icm.krasn.ru, msad@icm.krasn.ru

Введение

Одной из важнейших задач современной биологии является выявление и изучение структур в нуклеотидных последовательностях различного уровня и различного таксономического и функционального положения. Причем интерес вызывает как структурно-функциональный, так и эволюционный аспект. Эволюционное направление позволяет получить картину развития биологических систем от конкретного вида до глобальных сообществ. Выявление связи между структурой и функцией так же является классической задачей молекулярной и системной биологии и биоинформатики.

Настоящая работа посвящена изучению связи структуры нуклеотидных последовательностей и таксономии их носителей на примере геномов хлоропластов. Хлоропласты очень удобный объект для исследований, поскольку, во-первых, для них полностью исключается влияние различий в кодируемых функциях. Во-вторых, геномы хлоропластов имеют сравнительно малый размер ($\approx 10^5$ п.н.). В-третьих, происхождение хлоропластов традиционно ведут от бактерий [1]. Поток работ, посвящённых изучению эволюционных процессов и отношений в царстве растений на основе анализа особенностей структуры последовательностей геномов [2; 3] либо отдельных фрагментов и генов [4-6] хлоропластов, велик. При этом часто утверждения об эволюции вида делаются на основании анализа изменений генома хлоропластов.

Настоящая работа посвящена изучению структур в геномах хлоропластов и выявлению связи таких структур с филогенией их носителей. Под структурой мы будем понимать различие (либо подобие) статистических свойств отдельных формально выделяемых фрагментов генома на уровне триплетов. Иными словами, структура в рамках настоящей работы — это взаимное расположение различных (формально выделяемых) фрагментов генома сравнительно небольшой длины в пространстве частот триплетов, которые подсчитываются в пределах указанных фрагментов.

Ранее [7, 8] было показано, что геномы бактерий обладают выраженной семикластерной структурой. При разбиении генома бактерий на множество участков одинаковой длины (порядка $\approx 10^2$ нуклеотидов) и построении для каждого из них частотного словаря толщины 3 было обнаружено, что полученные словари кластеризуются в пространстве частот триплетов на семь кластеров. Данная работа посвящена проверке этого свойства для геномов хлоропластов различных наземных многоклеточных растений.

Рассматриваются генетические последовательности длины L , состоящие из символов алфавита $\mathcal{M} = \{A, C, G, T\}$. Для каждой из последовательностей составляется частотный словарь толщины 3. Под частотным словарем толщины 3 будем понимать список всех троек $v_1v_2v_3$ идущих подряд нуклеотидов с указанием частот этих троек. Всего может быть 64 триплета. Частота f_ω — это отношение числа копий n_ω данного слова к общему числу всех триплетов N , где N — сумма всех n_ω :

$$f_{\omega} = \frac{n_{\omega}}{N} \quad (1)$$

При этом словарь W_3 задает отображение генома в 64-мерное метрическое пространство. Два генома считаются близкими, если соответствующие им точки в 64-мерном пространстве близки в смысле Евклидовой метрики.

Проводилась предварительная обработка генетической последовательности. Последовательность сканировалась окном длины Δ с шагом t . Для каждого положения i рамки определялся участок генетической последовательности, совпадающий с рамкой считывания, для которого вычислялся частотный словарь $W_3^{(i)}$, соответствующий i -ой точке в 64-мерном пространстве. Кроме того, с каждой точкой в 64-мерном пространстве связывались следующие параметры: номер центрального символа рассматриваемого участка и относительная фаза.

Номер центрального символа участка совпадает с номером этого символа в последовательности. Относительная фаза определяется в зависимости от того, попал рассматриваемый участок в кодирующую или не кодирующую область последовательности. Участок относится к кодирующим, если он целиком попадал в кодирующую область последовательности. Если участок относится к не кодирующим, то соответствующая ему точка помечается символом J . Для кодирующего участка возможны 6 вариантов маркировки: $B_0, B_1, B_2, F_0, F_1, F_2$. Если кодирующий участок в последовательности аннотирован как считывающийся в прямом направлении, то для него вычислялся остаток от деления на 3 разности номеров центрального символа участка и первого символа кодирующей области, к которой он относится. В соответствии с величиной остатка от деления точка помечалась символом B_0, B_1 или B_2 . Если участок аннотирован как считывающийся в обратном направлении, то вычислялся остаток от деления на 3 разности номеров последнего символа кодирующей области, к которой относится участок, и центрального символа участка. В зависимости от значения остатка от деления точка помечалась символами F_0, F_1 или F_2 .

Данные для исследования брались в базе EMBL-банка. Были отобраны 185 хлоропластов наземных растений. Для всех генетических последовательностей длина рамки считывания $\Delta = 603$, шаг $t = 11$. По полученному множеству точек в программе *VidaExpert* (<http://bioinfo-out.curie.fr/projects/vidaexpert/>) строился вид данных в пространстве первых трёх главных компонент, вычисленных для данного 63-мерного пространства. Рассматривались две проекции на плоскость пространства главных компонент: вид данных “анфас” и “в профиль”. Точки данных были раскрашены в соответствии с их относительной фазой. Точки, соответствующие не кодирующим участкам, изображены на рисунках коричневым цветом, точки, соответствующие относительным фазам B_0 и F_0 , изображены темно малиновым и светло малиновым цветом, точки, соответствующие относительным фазам B_1 и F_1 , изображены темно зеленым и светло зеленым цветом, а точки, соответствующие относительным фазам B_2 и F_2 , изображены темно желтым и светло желтым цветом.

Кроме того, для каждого генома вычислялся GC-контент, то есть отношение суммарного числа нуклеотидов C и G к общему числу нуклеотидов в геноме.

Для облегчения восприятия результатов введём точные определения терминов, которые будут использоваться ниже. Выделяемые рамкой считывания длины Δ фрагменты помечались тремя разными метками: *ген*, *джанк* и *хвост*.

При этом исходно фрагменты получали одну из двух меток: *ген* или *джанк*.

- Фрагмент с номером i получал метку *ген*, если он попадал в ту часть генома, которая в аннотации, содержащейся в голове файла EMBL-банка, была помечена как CDS, либо как экзон, либо как tRNA, либо как rRNA, если длина такового превышала длину фрагмента. При этом фрагмент получал такую метку лишь при условии полного попадания в участок, аннотированный как описано выше.
- Фрагмент с номером i получал метку *джанк*, если он не получал метку *ген*, при условиях, описанных выше.
- Фрагмент с номером i получал метку *хвост* не на основании данных аннотирования генома, но на основании результатов кластеризации в пространстве главных компонент. Метку *хвост* получали те фрагменты, которые в пространстве трех первых главных компонент образовывали четко выраженный сравнительно небольшой кластер, лежащий на оси симметрии получающейся структуры. В *хвост* попадают фрагменты, относящиеся к *джанку*, tRNA и rRNA.

Результаты и обсуждение

Начнём представление результатов с описания базы исследованных геномов. Для каждого генома выделялись три типа фрагментов: *ген*, *джанк* и *хвост* (их детальное определение см. выше). Кроме того, мы определяли GC-контент для

- а) всего генома в целом,
- б) только для фрагментов, помеченных как *ген* (вне зависимости от того, в каком стрэнде они находятся),
- в) для фрагментов, помеченных как *джанк* и
- г) для фрагментов, располагающихся в «хвосте» (см. раздел 3.1).

Для 185 геномов хлоропластов были вычислены значения GC-контента, определяемого по всему геному в целом. У 87% геномов значение GC-контента лежит в интервале (0,35; 0,40).

Таблица 1. Коэффициенты корреляции между значениями GC-контента для разных фрагментов геномов хлоропластов и генома в целом.

	Ген	Джанк	Хвост
Геном	0,9745	0,9617	0,6248
Ген		0,9218	0,6285
Джанк			0,5939

У хлоропластов *Orthotrichum rogeri*, *Syntrichia ruralis*, *Physcomitrella patens*, *Marchantia polymorpha*, *Sanionia uncinata*, *Anthoceros angustus*, *Ptilidium pulcherrimum*, *Equisetum arvense*, *Glycyrrhiza glabra*, *Trifolium subterraneum*, *Orobanchе gracilis*, *Taxus mairei*, *Millettiapinnata*, *Pisum sativum*, *Juniperus virginiana* и *Juniperus bermudiana* значения GC-контента лежат ниже отметки в 0,35 (список приведён в порядке возрастания значений GC-контента). Противоположную группу составляют геномы видов *Aneura mirabilis*, *Lygodium japonicum*, *Pteridium aquilinum*, *Ophioglossum californicum*, *Marsilea crenata* и *Myriopteris lindheimeri*, у которых значения GC-контента превышают уровень в 0,40 (список также приведён в порядке возрастания значений); наконец, есть два генома-рекордсмена — *Selaginella moellendorffii* и *S. uncinata*, у которых значение GC-контента составляет 0,51003 и 0,54849, соответственно.

Значения GC-контента для фрагментов геномов, содержащих как гены, так и джанк, в целом близки к значениям GC-контента, определяемого для всего генома. Тот

факт, что согласованность в значениях GC-контента для генов и джанка со значениями GC-контента для всего генома весьма высока, следует из Таблицы 1, в которой представлены коэффициенты корреляции между значениями GC-контента для всего генома и для фрагментов, содержащих гены и джанк.

Было обнаружено, что все геномы «анфас» имеют выраженную трех лучевую структуру. Для подавляющего большинства геномов типичен вид «анфас» и «в профиль», который показан на рис. 1.

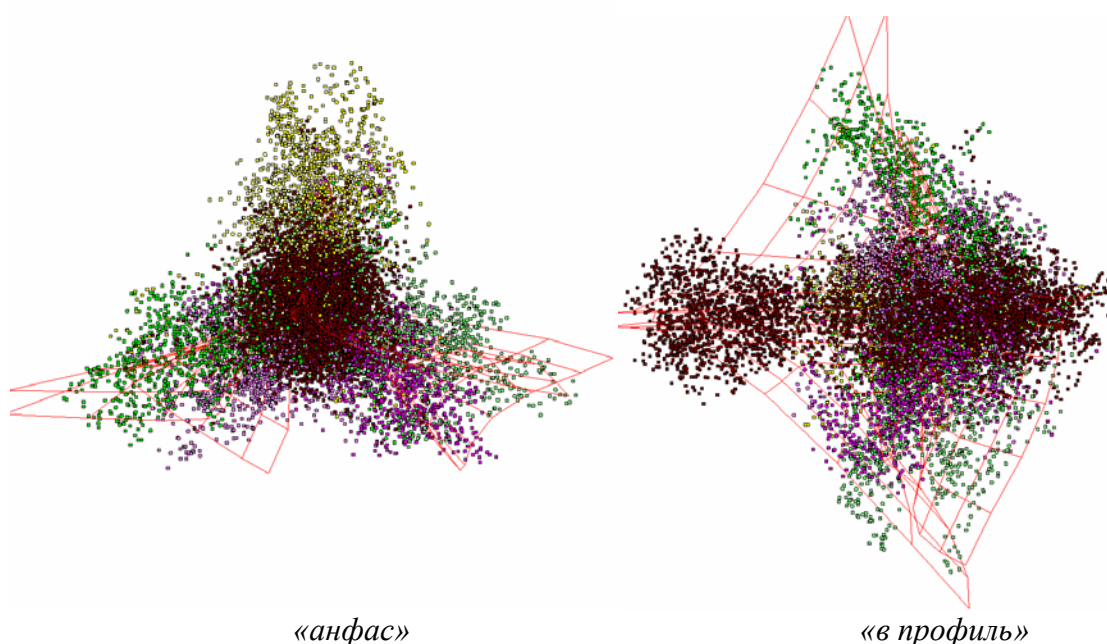


Рис. 1. Типичный вид структуры данных геномов наземных растений в проекциях пространства первых трех главных компонент (приведена структура генома *Nimphaea alba*).

Как видно на рис. 1, вид «анфас» хорошо кластеризован относительно точек из кодирующих и не кодирующих участков генома. Центральную часть рисунка «анфас» занимают точки из некодирующих участков, они окрашены в коричневый цвет. Один луч состоит из точек, относящихся к прямой и обратной относительной фазе 2 (точки, отмеченные светло желтым и темно желтым цветом), второй луч содержит точки, относящиеся к прямой относительной фазе 0 и обратной относительной фазе 1 (точки темно сиреневого и светло зеленого цвета), и третий состоит из точек, относящихся к обратной относительной фазе 0 и прямой относительной фазе 1 (точки светло сиреневого и темно зеленого цвета). На проекции “в профиль” (Рис. 2 а) видно, что точки, занимающие центральную часть распределения в проекции анфас, разделены на два кластера. Первый находится в центре трёхлучевой структуры, а второй обособлен от остальных точек и именно его мы и называем “хвостом”. Такой типичной структурой данных в пространстве первых трёх главных компонент обладают геномы 179 хлоропластов из 185 рассматриваемых. Собственно, наличие этого разбиения фрагментов на два кластера и позволяет утверждать существование восьмикластерной, а не семикластерной (как в работах (Gorban et al., 20036, 2005))

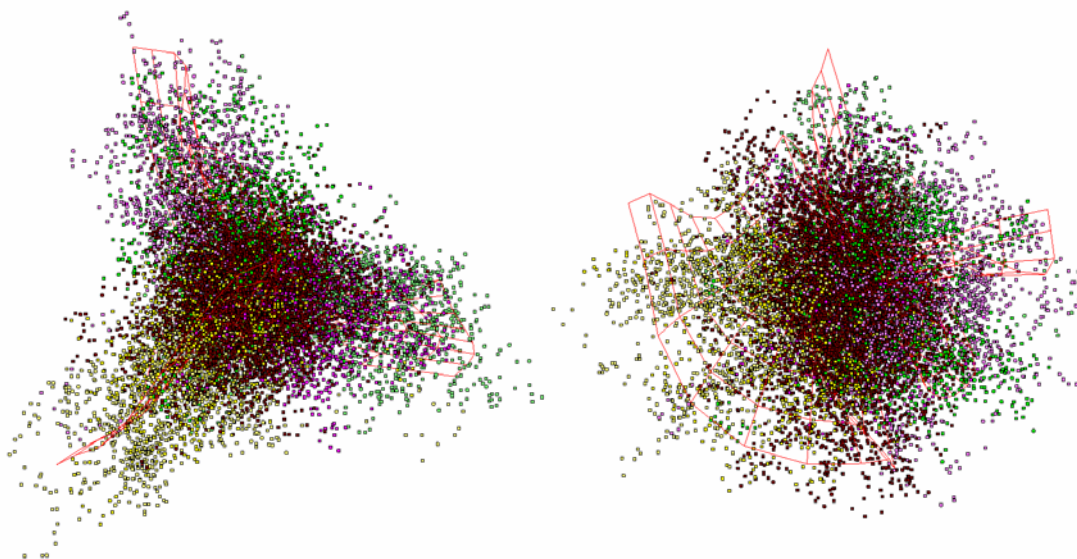


Рис. 2. Структура данных в проекциях пространства трех первых главных компонент *Selaginella moellendorffii*.

У оставшихся геномов хлоропластов наблюдались следующие отличия от типичной структуры данных. У геномов хлоропластов *Selaginella moellendorffii* и *Selaginella uncinata* отсутствует обособленная группа точек, относящаяся к некодирующим участкам генома. Это представители древних споровых растений, которые впервые появились около 400 млн. лет назад. Структура данных этих геномов существенно больше похожа на структуру данных геномов бактерий (см. Рис.2).

Геном хлоропласта *Anthoceros angustus* отличается от структуры, характерной для подавляющего большинства геномов хлоропластов: здесь обособленный участок данных этого хлоропласта состоит из точек белок- кодирующих участков (рис. 3). Это единственный представитель *Anthoceros* в рассматриваемой базе генетических данных.

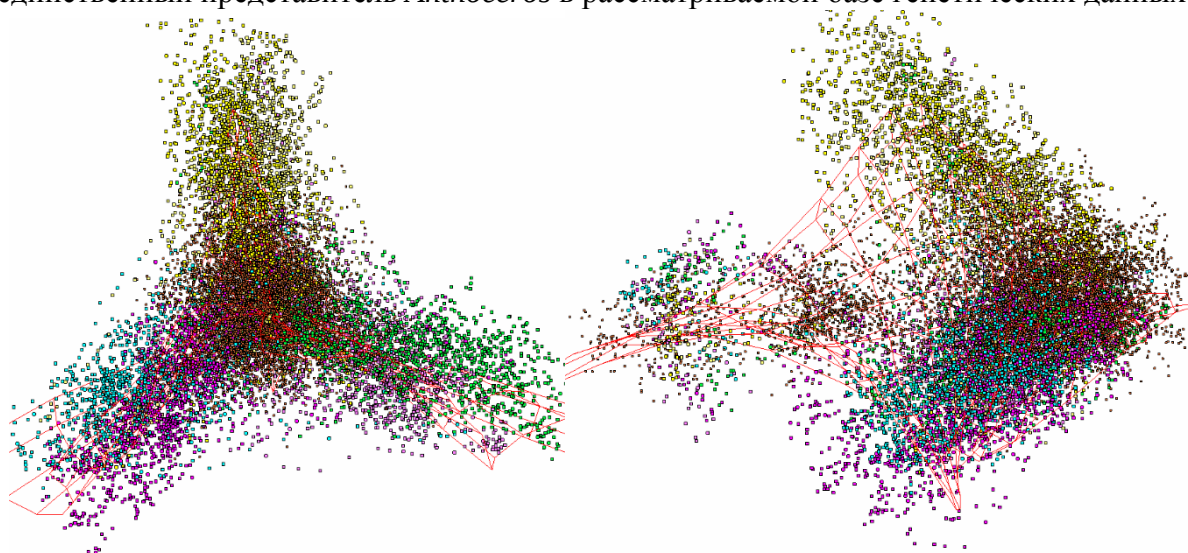


Рис. 3. Структура данных в проекциях пространства трех первых главных компонент *Anthoceros angustus*.

У растений *Brachypodium distachyon*, *Juniperus bermudiana*, *Pinus contorta*, *Jacobaea vulgaris*, *Najas flexilis*, относительные фазы в лучах перемешаны (см. рис. 4).

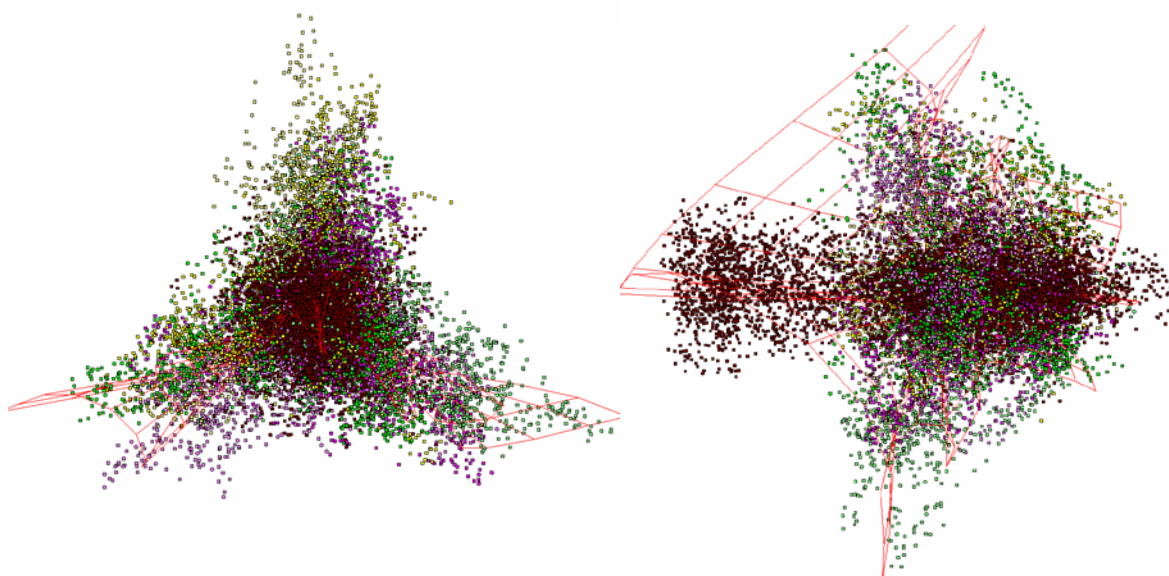


Рис. 4. Структура данных со смешанными фазами в лучах (*Brachypodium distachyon*).

Кроме того, обнаружилось, что для мхов CG-контент меньше 0.3 и изменялся от 0.28361 у *Orthotrichum rogeri* до 0.293164 у *Sanionia uncinata*. Для папоротников CG-контент больше 0.4 и изменялся от 0.406422 у *Lygodium japonicum* до 0.427245 у *Myriopteris lindheimeri*. У *Selaginella moellendorffii* и *Selaginella uncinata*, представителей древних споровых высших растений, CG-контент составил 0.510026 и 0.548491, соответственно. Для всех остальных геномов хлоропластов CG-контент изменялся от 0.3 до 0.4.

Таким образом, для геномов хлоропластов с GC-контентом ниже 50% не установлено влияния GC-контента на структуру данных, однако этот показатель сильно коррелирует с филогенией.

Кроме кластеризации фрагментов, были построены графики изменения GC-контента вдоль по генетической последовательности по этим фрагментам. Для этого вычислялся GC-контент каждого фрагмента последовательности, которые использовались для построения кластеризации (напомним, длина фрагмента $\Delta = 603$, сдвиг $t = 11$). На Рис. 5 показан типичный график изменения GC-контента для хлоропластов. Как видно из рисунка, график имеет во второй половине два пика с двумя зубцами, значения GC-контента в которых превышает 0,5. Зубцы с высокой точностью соответствуют локализации фрагментов из “хвоста” в генетической последовательности. Такого вида график характерен для 162 последовательностей хлоропластов из 185. У 12 графиков во второй половине находится только один пик с двумя зубцами (такого вида график у *Nageia nagi*, *Trifolium subterraneum*, *Pinus contorta*, *Picea abies*, *Lathyrus sativus*, *Hordeum vulgare*, *Triticum aestivum*, *Glycyrrhiza glabra*, *Podocarpus lambertii*, *Retrophyllum piresii*, *Pseudotsuga sinensis*, *Pinus taeda*. *Hordeum vulgare* и *Triticum aestivum* попали в этот список по следующим, скорее формальным, причинам. В базах данных EMBL и GeneBank последовательность *Hordeum vulgare* ошибочно депонирована как полная, хотя она не содержит инвертированного повтора. Для *Triticum aestivum* в EMBL содержится две аннотации разных авторов. Одна с инвертированным повтором (AB042240), другая без него (KC912694). И для оставшихся 11 графиков нет какого-то типичного вида: встречаются два пика в начале, один пик в первой половине

последовательности, один пик в начале и один пик в конце и т.д. Этим графикам соответствуют хлоропласты следующих растений: *Keteleeria davidiana*, *Pelargonium × hortorum*, *Pisum sativum*, *Stangeria eriopus*, *Juniperus bermudiana*, *Juniperus monosperma*, *Juniperus virginiana*, *Taxus mairei*, *Larix decidua*, *Cycas revoluta*, *Juniperus bermudiana*. Для всех изученных последовательностей, зубцы пиков совпадают с локализацией 16S и 23S rRNA. На положение пиков на графиках влияет расположение этих rRNA в геноме. Различие графиков обусловлено различным расположением rRNA в генетических последовательностях. Иных функционально нагруженных участков в этих фрагментах не обнаружено, по крайней мере, для той аннотации геномов, которая нами использовалась.

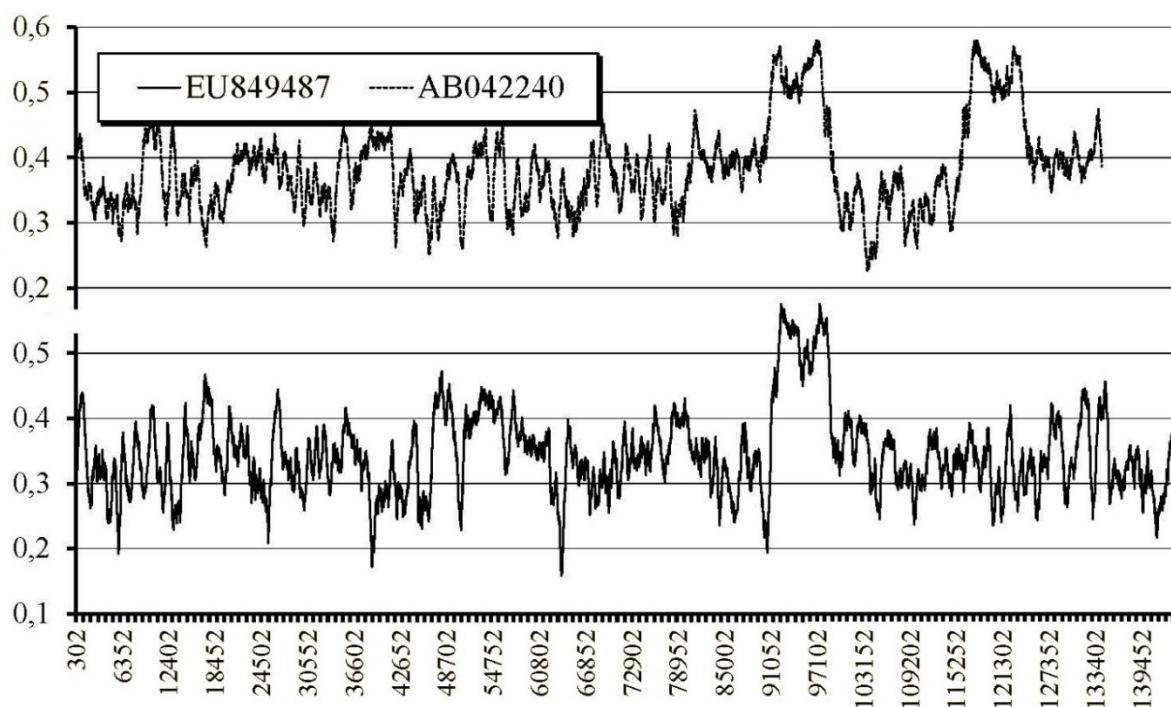


Рис. 7. Типичный график изменения GC-контента геномов хлоропластов на примере *Triticum aestivum*(AB042240) и *Trifolium subterraneum*(EU849487)

Было обнаружено, что джанк геномов хлоропластов в проекциях первых трех главных компонент 63-мерного пространства частотных словарей компактно занимает одну и ту же область. На рис.8 в качестве примера показан джанк следующих геномов: *Physcomitrella patens* (точки красного цвета), *Syntrichia ruralis* (точки синего цвета), *Ptilidium pulcherrimum* (точки зеленого цвета), *Sanionia uncinata* (точки желтого цвета), *Orthotrichum rogeri* (точки малинового цвета).

Однако джанк геномов *Selaginella moellendorffii* и *Selaginella uncinata* выделяется из общей картины. На рис.9 видно, что он расположен достаточно обособленно по отношению к джанку остальных геномов. На рис.9 показан джанк следующих геномов: *Triticum aestivum* (точки красного цвета) и *Selaginella moellendorffii* и *Selaginella uncinata* (точки синего и зеленого цвета соответственно). Здесь точки, соответствующие *Triticum aestivum* занимают типичную область расположения, а точки, соответствующие *Selaginella moellendorffii* и *Selaginella uncinata* сдвинуты относительно типичного расположения.

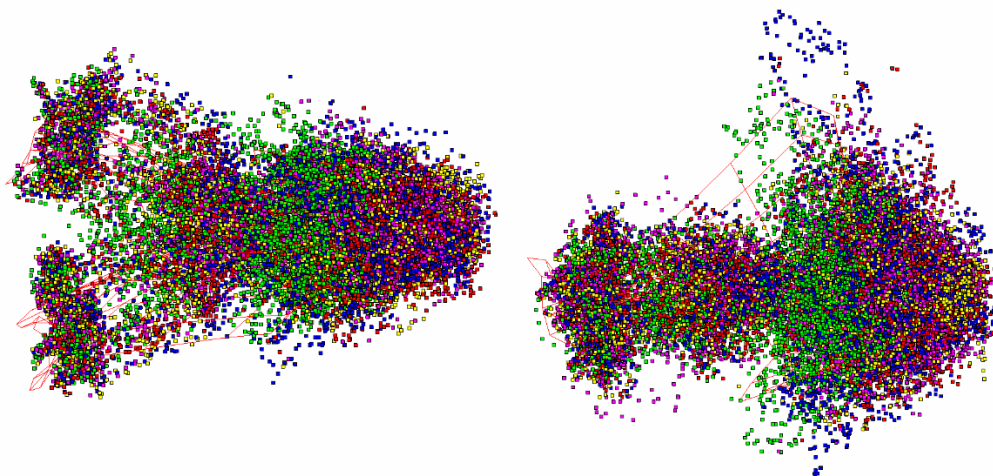


Рис 8. Джанк геномов *Physcomitrella patens*, *Syntrichia ruralis*, *Ptilidium pulcherrimum*, *Sanionia uncinata*, *Orthotrichum rogeri* вид «анфас» и в профиль в пространстве первых трех главных компонент.

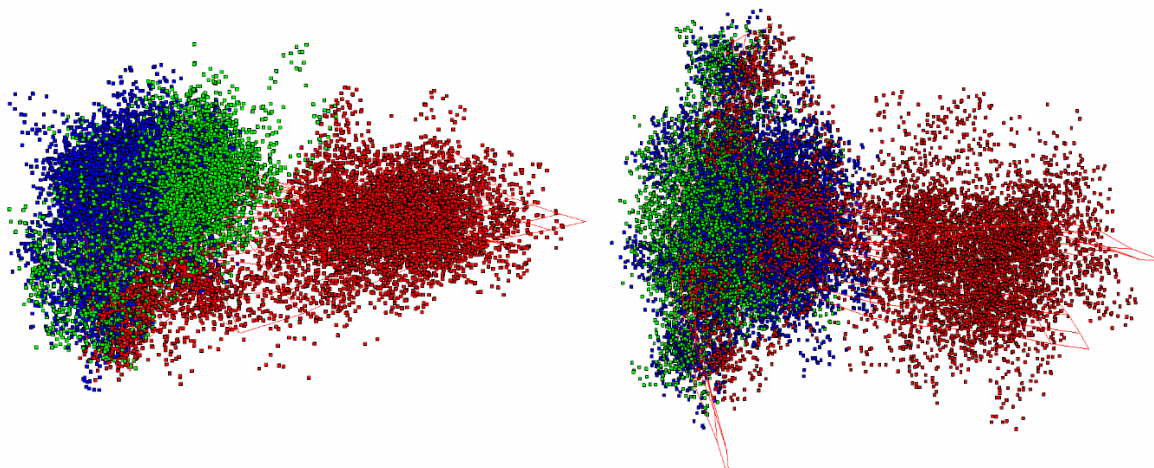


Рис 8. Джанк геномов *Physcomitrella patens*, *Syntrichia ruralis*, *Ptilidium pulcherrimum*, *Sanionia uncinata*, *Orthotrichum rogeri* вид «анфас» и в профиль в пространстве первых трех главных компонент.

Заключение

Для большинства рассмотренных геномов хлоропластов наземных растений установлено, что структура данных в пространстве первых трех главных компонент очень похожа: имеется центральное ядро из фрагментов некодирующих участков и три луча, состоящих из фрагментов кодирующих участков. Кроме того, имеется отдельная группа фрагментов некодирующих участков, не входящая в центральное ядро. Таким образом, можно говорить о воямикластерной структуре геномов хлоропластов. Следует заметить, что геномы хлоропластов обладают заметными отличиями по своей структуре от геномов бактерий: у них выделяется группа фрагментов, соответствующих некодирующим областям и областям *tRNA* и *rRNA*, которые в пространстве главных компонент выглядят как своего рода «хвост». Этой особенности не наблюдается у геномов бактерий. Проведённые исследования показывают, что в отличие от бактериальных геномов, CG-контент не является здесь ведущим фактором

распределения на типы геномов по числу кластеров и фактически все геномы обладают восьмикластерной структурой. Функциональная роль и эволюционные механизмы возникновения таких групп фрагментов требуют специального изучения.

Список литературы

1. Мережковский К.С. 1909. Теория двух плазм как основа симбиогенезиса, нового учения о происхождении организмов. Казань: Изд-во Импер. ун-та, 102 с.
2. Leliaert F., Smith D.R., Moreau H., Herron M.D., Verbruggen H., Delwiche Ch.F., De Clerck O., 2012. Phylogeny and Molecular Evolution of the Green Algae // *Critical Reviews in Plant Sciences*. V. 31. P. 1-46.
3. Carbonell-Caballero J., Alonso R., Ibanez V., Terol J., Talon M., Dopazo J., 2015. A phylogenetic analysis of 34 chloroplast genomes elucidates the relationships between wild and domestic species within the genus citrus // *Molecular Biology and Evolution*. V. 32. № 8. P. 2015--2035.
4. Milanowski R., Zakrys B., Kwiatowski J., 2001. Phylogenetic analysis of chloroplast smallsubunit rRNA genes of the genus *Euglena* Ehrenberg // *International Journal of Systematic and Evolutionary Microbiology*. V. 51. P. 773-781.
5. Dong W. , Liu J., Yu J., Wang L., Zhou Sh., 2012. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding // *PLoS ONE*. V. 7. № 4. P. 1-9.
6. Marazzi B., Endress P.K., De Queiroz L.P., Conti E., 2006. Phylogenetic relationships within senna (leguminosae, cassiinae) based on three chloroplast dna regions: patterns in the evolution of floral symmetry and extrafloral nectaries // *American Journal of Botany*. V. 93. № 2. P. 288-303.
7. Gorban A. N., Zinovyev A. Yu., Popova T. G. (2003) Seven clusters in genomic triplet distributions. In *Silico Biology* 3: 39-45.
8. Gorban A. N., Zinovyev A. Yu., Popova T. G. (2005) Universal seven-cluster structure of genome fragment distribution: basic symmetry in triplet frequencies. *Bioinformatics of Genome Regulation and Structure II*. (Eds. N. Kolchanov and R. Hofstaedt) Springer Science+Business Media, Inc. P. 153-163.
9. McFadden G. I. (2001) Chloroplast Origin and Integration // *Plant Physiology*. 125(1), p. 50-53.
10. Bugaenko N. N., Gorban A.N., Sadovsky M. G. (1998) Maximum entropy method in analysis of genetic text and measurement of its information content. *Open Systems & Information Dyn.* 5(2): 265-278.
11. Gorban A.N., Zinovyev A.,Yu. Principal manifolds and graphs in practice: from molecular biology to dynamical systems // *Int. J. of Neural Systems*, 2010, Vol. 20, № 3, pp. 219 – 232.

ТЕОРИИ ВСЕГО И СМЫСЛОВОЕ РАЗВИТИЕ

В.П.Сизиков¹, В.И. Разумов²

¹Омский государственный университет путей сообщения, v_p_sizikov@mail.ru

²Омский государственный университет им. Ф.М. Достоевского, razumovvi@omsu.ru

1. Введение. В работе [1] актуализировалось понятие теории всего ($\wedge$ ТВ). И на данный момент сформированы качественные модели уже четырёх $\wedge$ ТВ: рода базовых дисциплин, модернизации как стратегии развития, рода измерительной процедуры [1], рода знания как стратегии познания. А в работе [2] осуществлён системный анализ функционирования динамической информационной системы (ДИС) на примере $\wedge$ ТВ, описывающей модернизации, с серией выводов на тему развития общества.

Накопленный материал позволяет признать $\wedge$ ТВ фундаментальным понятием, тесно увязать его с понятием развития и позаботиться о формировании соответствующей математической базы. Здесь дана база на уровне качественных моделей $\wedge$ ТВ.

2. Инструментарий формирования объективных законов для теорий всего. Считаем, что вся система знаний в ранге номологической базы (НЛ-Б') [3] представлена ДИС-компьютером (ДИС-*К) Γ_0 уровня 4, имеющим 81 вершину как универсальные мини-категории. Формально, любая мутация у Γ_0 , представленная развёрткой в четыре шага дешифровки как режимом процесса рассуждения, выступает как некая проекция мировоззрения. Однако в [3] были выявлены 27 триад, составляющих вместе Γ_0 , получивших статус *законов* в НЛ-Б'. Так что наиболее осмысленными и онтологически значимыми оказываются те мутации у Γ_0 , в которых режим процесса рассуждения укладывается в серию из 27 законов в НЛ-Б'. Так дело сводится к выбору конкретной развёртки G ДИС-*К Γ_0 уровня 3, вершинами которого выступают эти 27 законов.

За развёрткой G из 27 законов закрепляется группа из шести её *базовых мутаций* [4], среди них две ориентированы на философские, две – на физические и две – на математические начала. Осмысленные комбинации на 27 законах, образующие триады в базовых мутациях, понимаются как *объективные законы* той конкретной реальности, что скрывается за развёрткой G . А комбинации, образующие ДИС-*К уровня 2, понимаются как *представители* этой реальности. В мутациях, ориентированных на философские начала, объективные законы выступают в статусе *механизмов*, на физические начала – *качеств*, на математические начала – *ролей*.

В каждой из развёрток можно осуществить дешифровки законов до мини-категорий в НЛ-Б' и произвести структурный анализ с осмыслением на более тонком уровне. Однако это актуально уже после описания на языке свода законов.

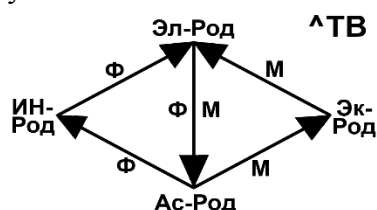


Рис. 1. Рода теории всего ($\wedge$ ТВ): Ас- – аспектный, ИН- – инструментальный, М – математика, Ф – физика, Эк- – экспериментальный, Эл- – элементный.

И главное, что теперь к развёртке G ДИС-*К G_0^* уровня 3 добавляется группа из её, так называемых, *родовых мутаций*. Вместе они и образуют $\wedge$ ТВ из четырёх её *родов* – *аспектный, инструментальный, экспериментальный и элементный* род (Ас-, ИН-, Эк- и Эл-Род). Причём эти рода увязаны схемой когнитивной ячейки (рис. 1).

В принципе, исходная развёртка G может явиться на месте любого из этих родов, требуя доопределения остальных родов $\wedge$ ТВ. Однако последнюю процедуру, да и суть самих родовых мутаций, проще понять, когда G приходится на место Ас-Рода.

Итак, пусть развёртка G как ДИС-*К уровня 3 из 27 законов представляет Ас-Род $\wedge$ ТВ с уже определившимися в нём механизмами, качествами, ролями, объективными законами и представителями. На этом ДИС-*К вводится родовая мутация физической направленности как суперпозиции двух преобразований. Первое из них – обычная базовая мутация физической направленности на каждом из представителей в G [4], приводящая к качествам вместо механизмов. Второе – такая же мутация, но на наборе самих качеств как категорий. В итоге получится новый ДИС-*К G_f уровня 3 из 27 законов, представляющий родовую мутацию физической направленности для G . Это будет ИН-Род $\wedge$ ТВ. И сразу ясно, что механизмы ИН-Рода – это качества Ас-Рода, но в новых комбинациях, выражающих совершенно других представителей $\wedge$ ТВ.

Аналогично на ДИС-*К G вводится родовая мутация математической направленности как суперпозиции двух преобразований, действующих по образцу обычной базовой мутации математической направленности. Результат G_m её действия на G представит уже Эк-Род $\wedge$ ТВ. И сразу ясно, что механизмы Эк-Рода – это роли Ас-Рода, но в новых комбинациях, выражающих совершенно других представителей $\wedge$ ТВ с новыми у них качествами и ролями, за исключением, однако, одного момента. Дело в том, что механизмы, качества и роли в каждом ДИС-*К уровня 3 тоже связаны базовыми мутациями. В итоге оказывается, что качества в ИН-Роде и роли в Эк-Роде тоже одинаковы, но в разных комбинациях, т.е. новизна у Эк-Рода останется лишь за качествами.

Если к G_f как ДИС-*К уровня 3 из 27 законов применить ещё раз родовую мутацию физической направленности, то получится в итоге результат G_{ff} , который представит уже Эл-Род $\wedge$ ТВ. И сразу видно, что механизмы Эл-Рода – это тоже качества ИН-Рода, но опять в новых комбинациях, выражающих совершенно других представителей $\wedge$ ТВ с новыми у них качествами и ролями, за исключением вновь одного момента. Дело в том, что применение к G_{ff} родовой мутации физической направленности даёт исходную развёртку G , представляющую Ас-Род $\wedge$ ТВ. Так что механизмы Ас-Рода и качества Эл-Рода повторяют друг друга, новизна у Эл-Рода останется лишь за ролями.

Также можно к G_m как ДИС-*К уровня 3 из 27 законов применить ещё раз родовую мутацию математической направленности, получив в итоге результат G_{mm} , у которого механизмы повторяют роли Эк-Рода (они же и качества ИН-Рода), а роли – механизмы Ас-Рода, так как применение к G_{mm} родовой мутации математической направленности тоже даёт исходную развёртку G , представляющую Ас-Род $\wedge$ ТВ. К тому же, качества у результата G_{mm} повторяют роли Эл-Рода и, наоборот. Более того, представители у результата G_{mm} окажутся те же, что и у Эл-Рода, но порядок их расположения заменится на обратный. Все объективные законы и представители в G_{mm} повторяют таковые из G_{ff} , но с инверсией при этом статусов качеств и ролей. В русле $\wedge$ ТВ результат G_{mm} тоже следует считать представляющим её Эл-Род, хотя кое-где разницу между G_{mm} и G_{ff} следует всё же учитывать подобно разнице между становлением и приложением.

Более детально связи между родами произвольной $\wedge$ ТВ через структуру родов приведены на рис. 2–4, используя каноническую нумерацию вершин ДИС-*К уровня 3, представляющих 27 законов. На рис. 2 приведён Ас-Род с его механизмами, качествами и ролями, на рис. 3 – ИН-Род и Эк-Род, а на рис. 4 – Эл-Род в двух вариантах, соответствующих родовым мутациям физической и математической направленности.

Обратим ещё раз внимание на то, что некоторые объективные законы используются сразу в нескольких родах $\wedge$ ТВ. При этом Ас-Род выступает как точка ветвления между остальными родами, представляющими вместе триединство воплощения реальности.

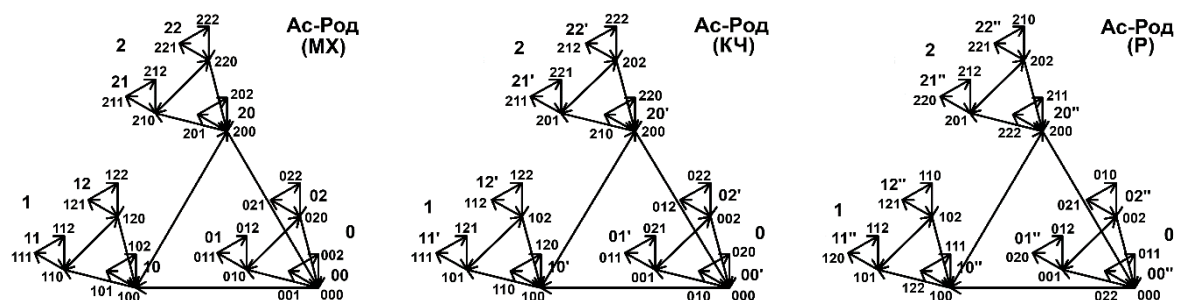


Рис. 2. Структура аспектного рода теории всего. Обозначения, дополняющие таковые с рис. 1: КЧ – качества, МХ – механизмы, Р – роли.

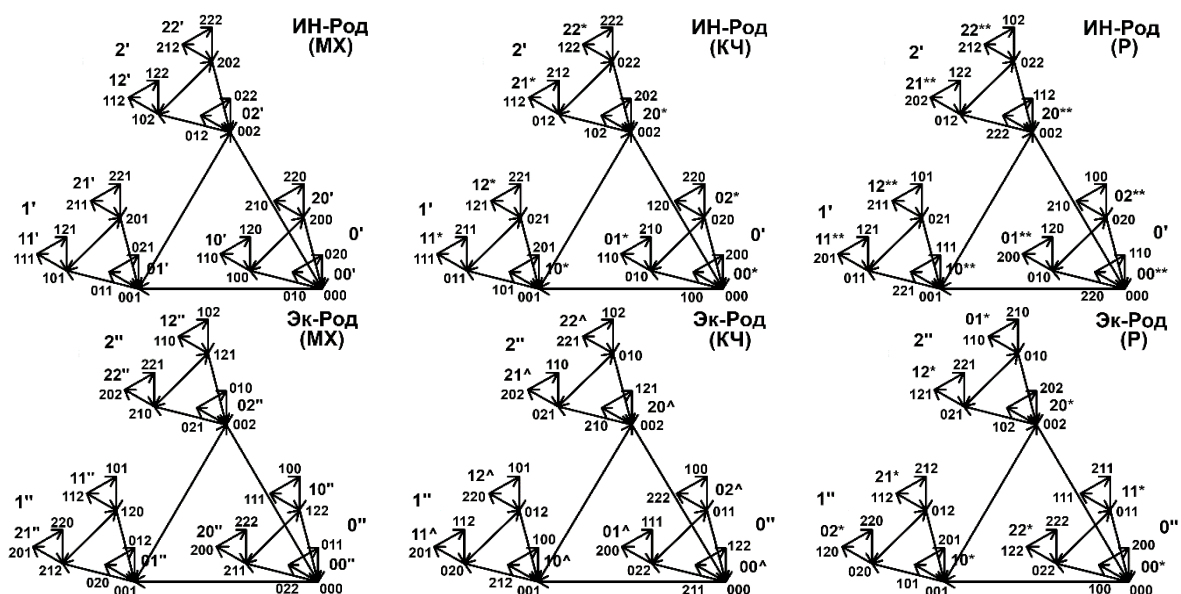
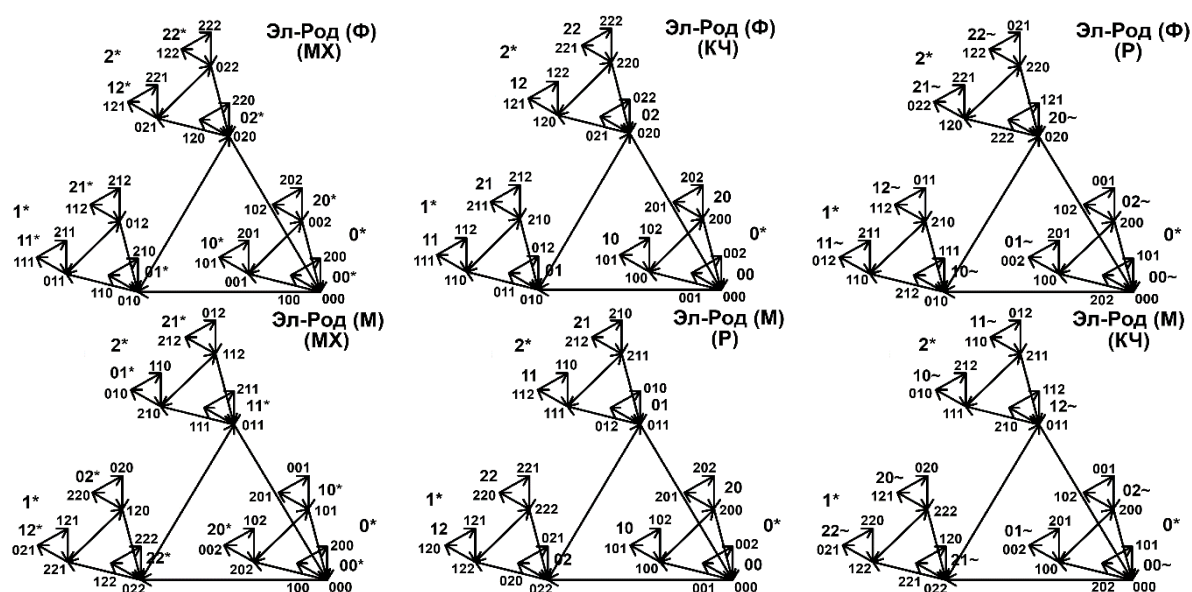


Рис. 3. Структура инструментального и экспериментального родов теории всего. Обозначения с рис. 1, 2.



Определение теории всего. Под $\wedge\text{ТВ}$ понимается произвольная согласованная группа приведённых на рис. 2–4 схем четырёх родов этой $\wedge\text{ТВ}$, в которых вершины с трёхзначными номерами – 27 законов из НЛ-Б'.

Теорема 1. *Каждая $\wedge\text{ТВ}$ содержит 4 рода, 12 представителей, 63 объективных закона и 27 простых законов. При этом рода связаны схемой когнитивной ячейки, осью которой является переход от Эл-Рода к Ас-Роду (рис. 1).*

3. Классы сопряжённых теорий всего. Формирование всей структурной базы – *качественной модели* конкретной $\wedge\text{ТВ}$ производится с использованием двух родовых мутаций – физической и математической направленности по схеме на рис. 1 или через согласование схем на рис. 2–4. В ранге исходного может выступить любой из четырёх родов $\wedge\text{ТВ}$. И всё будет восстанавливаться однозначно, за исключением случая, когда исходным берётся Эл-Род. Если Эл-Род выступит в варианте G_{ff} , надо по нему формировать Ас-Род G с помощью родовой мутации физической направленности или используя левую часть на рис. 4. А если Эл-Род выступит в варианте G_{mm} , надо по нему формировать Ас-Род G с помощью родовой мутации математической направленности или используя правую часть на рис. 4. Итогом будут две разных $\wedge\text{ТВ}$, имеющих 6 общих и 6 различающих их представителей, 45 общих и 18 различающих их объективных законов, да и порядок расположения в схемах представителей и объективных законов изменится. Так возникает понятие сопряжённых $\wedge\text{ТВ}$.

Определение сопряжённых теорий всего. *Сопряжёнными* называются такие $\wedge\text{ТВ}$, у которых в Эл-Рode одинаковы представители и объективные законы в них, возможна разве что инверсия между законами-качествами и законами-ролями.

Учтём, что Эл-Род имеет три представителя с тремя объективными законами-механизмами в каждом, причём представитель и объективный закон, содержащие начало отсчёта, помеченные вершиной с номером 000, закреплены, а два других представителя и остальные законы в них могут меняться местами. И это позволяет вычислить количество элементов в такой совокупности.

Теорема 2. *Набор всех $\wedge\text{ТВ}$ распадается на непересекающиеся классы сопряжённых $\wedge\text{ТВ}$ по 144 различных сопряжённых $\wedge\text{ТВ}$ в каждом.*

Пусть H – один из классов сопряжённых $\wedge\text{ТВ}$, а H^* – все возможные в нём схемы Эл-Рода. Тогда при каждой $\wedge\text{ТВ}$ $h \in H$ имеется пара (h_f^*, h_m^*) схем из H^* , первая из которых – вариант G_{ff} по отношению к Ас-Роду G у h , а вторая – вариант G_{mm} . При этом h_f^* и h_m^* однозначно определяют друг друга. Условимся пару (h_f^*, h_m^*) выделять направленной связью от h_f^* к h_m^* . Это даст орграф G_H , вершинами которого служат элементы из H^* и их количество равно 144. Назовём этот орграф *базовым* для класса H .

Теорема 3. *У всех классов H сопряжённых $\wedge\text{ТВ}$ их базовые орграфы G_H изоморфны и представляют объединение из 24 изолированных циклов длины 6.*

Любой цикл длины 6 из базового орграфа G_H представлен на рис. 5 последовательностью (считая слева направо и сверху вниз) ДИС-*К уровня 2, каждый из которых выражает порядок объективных законов-механизмов в Эл-Рode. А на рис. 6 этот же цикл представлен через рода входящих в него $\wedge\text{ТВ}$. Назовём такие циклы *базовыми*, $\wedge\text{ТВ}$, имеющие Эл-Рода из одного базового цикла, *внутренне сопряжёнными*, а $\wedge\text{ТВ}$, у которых Эл-Рода принадлежат разным базовым циклам, *внешне сопряжёнными*.

В свою очередь, есть такие перестановки в первой схеме цикла, которые порождают различные базовые циклы, приумножая в итоге их количество. Это проиллюстрировано на рис. 7. Первая строка связана с круговой перестановкой элементов в представителе 1^* без изменений в представителях 0^* и 2^* , что даёт трехкратное приумножение количества базовых циклов. Во второй строке приведены три независимых варианта перемен в первой схеме, каждый из которых даёт двукратное

приумножение количества циклов: смена ориентации расположения элементов в представителе 1^* без изменений в представителях 0^* и 2^* ; то же в представителе 0^* без изменений в представителях 1^* и 2^* ; перестановка представителей 1^* и 2^* без изменений в самих представителях. Так в итоге получается 24 изолированных друг от друга базовых цикла в одном классе сопряжённых $\wedge TB$.

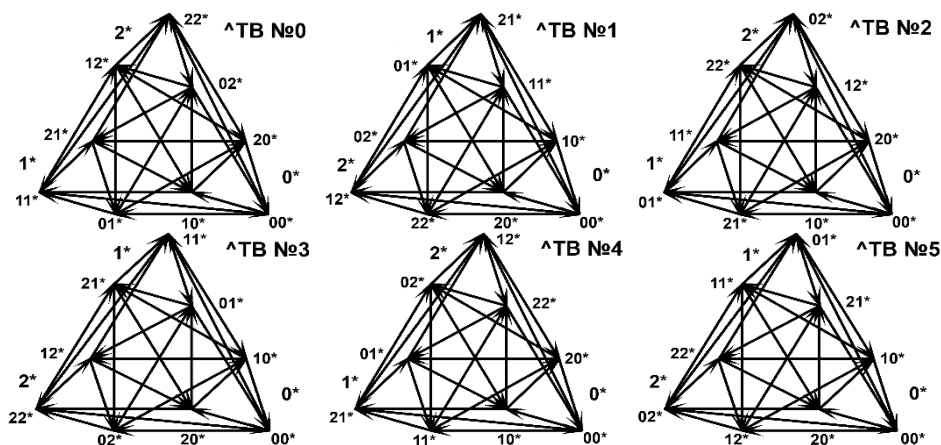


Рис. 5. Базовый цикл из теорий всего. Обозначения с рис. 1, 4.

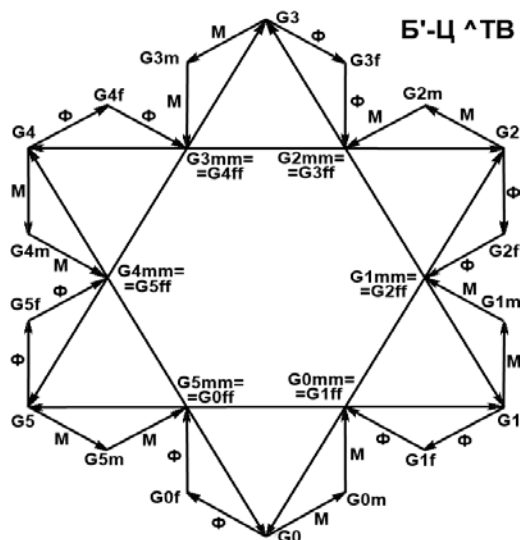


Рис. 6. Схема базового цикла из теорий всего (Б'-Ц $\wedge TB$): М – математика, Φ – физика; G_i ($i=0,1,2,3,4,5$) – Ас-Род, G_{if} – Ин-Род, G_{im} – Эк-Род, G_{iff} и G_{imm} – Эл-Род, полученный из G_i через физическую и математическую направленность.

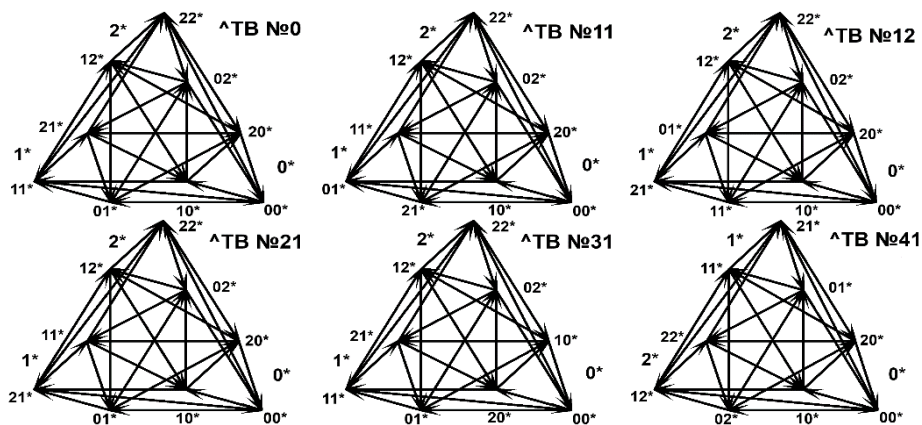


Рис. 7. Варианты приумножения количества базовых циклов

среди сопряжённых теорий всего. Обозначения с рис. 1, 4.

Итак, простейшими конструкциями по единению $\wedge$ ТВ выступают базовые циклы, включающие по 6 внутренне сопряжённых $\wedge$ ТВ. Конечно, эти $\wedge$ ТВ во многом повторяют друг друга и всё же содержат качественные различия.

4. К развитию через теории всего на уровне смыслов. Сведения из п. 2–3 делают актуальными варианты развития в $\wedge$ ТВ на уровне смыслов. И здесь выделяются варианты внешнего, внутреннего и трансформирующего развития.

Под *внешним* развитием в $\wedge$ ТВ на уровне смыслов понимается развёртка этой $\wedge$ ТВ до класса всех сопряжённых ей $\wedge$ ТВ. При этом базовые циклы из сопряжённых $\wedge$ ТВ называются *смысловыми единицами развития*. Основу дают сведения на рис. 6 и 7.

Под *внутренним* развитием в $\wedge$ ТВ на уровне смыслов понимается выстраивание базовых мутаций у каждого её представителя, предварительно дешифрованного до мини-категорий НЛ-Б' и в механизмах, и в качествах, и в ролях. Получающиеся при этом в схемах мутаций новые осмысленные триады из мини-категорий НЛ-Б' называются *производными законами* $\wedge$ ТВ. Это полностью осуществлено для всех перечисленных выше $\wedge$ ТВ, и результаты этого хранятся на сайте <http://thoughttring.com/ViewForm.aspx?id=421>.

Под *трансформирующим* развитием в $\wedge$ ТВ на уровне смыслов понимается использование одинаковых объективных законов в разных $\wedge$ ТВ и их родах для переноса смысловых ассоциаций, родившихся в одном месте, на другие места. При этом в случаях использования одинаковых объективных законов в разных родах одной $\wedge$ ТВ трансформирующее развитие называется *балансовым*.

Указанные варианты развития $\wedge$ ТВ на уровне смыслов явно свидетельствуют о наличии взаимодействия между $\wedge$ ТВ и их родами, в котором в роли ресурса выступают смыслы. И исследования в [2] использовали это на функциональном уровне.

5. Заключение. В работе предложен подход к формированию моделей, ориентированный на выражение развития и экспликацию смыслов. Обращение к $\wedge$ ТВ представляет опыт конструирования законов, двигаясь от знаний, организованных на основе ДИС-технологии в НЛ-Б'. Схемы, включенные в материал, являются в т.ч. шаблонами для организации знаний любого типа.

Список литературы

1. Сизиков В.П. Измерительные процедуры и теории всего / В.П. Сизиков, В.И. Разумов // Современные тенденции развития науки и технологий: Сб. науч. тр. по матер. XXV Междун. науч.-прак. конф. 29 апреля 2017 г.: в 5 ч. / Под общ. ред. Ж.А. Шаповал. – Белгород: ООО АПНИ, 2017. – № 4. Ч. I. – С. 23–36. То же [Электронный ресурс]. Режим доступа: <http://www.issledo.ru>.
2. Сизиков В.П. Системный анализ стратегий развития общества / В.П. Сизиков, В.И. Разумов // Современные тенденции развития науки и технологий: Сб. науч. тр. по матер. XXIV Междун. науч.-прак. конф. 31 марта 2017 г. – Белгород: ИП Ткачёва Е.П., 2017. – № 3–11. – С. 14–23. То же [Электронный ресурс]. Режим доступа: <http://www.issledo.ru>.
3. Разумов В.И. Приложения информационных основ синтеза систем: ДИС-развёртки и парадигма закона. Монография [Электронный ресурс] / В.И. Разумов, В.П. Сизиков. Электронное издание № 33934. № гос. регистр. 0321304636, ФГУП НТЦ «Информрегистр». – Омск. 2014.
4. Разумов В.И. Автоматизация рассуждений: программирование мутаций ДИС-компьютера уровня 2 / В.И. Разумов, В.П. Сизиков // Сибирский философский журнал. – 2016. – Т. 14. – № 1. – С. 53–69.

ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОЦЕНКИ ПОДГОТОВКИ СТУДЕНТОВ К ВЫПОЛНЕНИЮ ТРУДОВЫХ ФУНКЦИЙ ПРОФЕССИОНАЛЬНЫХ СТАНДАРТОВ

Н.В. Смыкова¹, О.И. Пятковский¹, Н.А. Звеков¹

¹Алтайский государственный технический университет им. И.И. Ползунова,
кафедра информационных систем в экономике, horyzont1@mail.ru

В настоящее время становится актуальной задача определения соответствия качества обучения профессиональным стандартам. 1 июля 2016 года вступил в силу 122-ФЗ, вносящий изменения в Трудовой кодекс РФ и закон «Об образовании», в частности формирование требований ФГОС к результатам освоения образовательных программ в части профессиональных компетенций будет осуществляться на основе соответствующих профессиональных стандартов [3]. Сейчас ведется разработка ФГОС ВО нового поколения, после введение в действие которых образовательные организации будут обязаны вести образовательную деятельность с учетом профессиональных стандартов.

Трудовой кодекс РФ определяет профессиональный стандарт, как характеристику квалификации, необходимой работнику для осуществления определенного вида профессиональной деятельности [2]. То есть по сути в профессиональном стандарте отражается результат обучения. В настоящий момент результат освоения образовательной программы выражается в освоении общекультурных, общепрофессиональных и профессиональных компетенций.

Если же говорить в целом об оценке профпригодности выпускников вуза, то необходимо оценивать не только соответствие их подготовки профессиональным стандартам, но учитывать личностные характеристики, такие как тип мышления, темперамент, склад ума и прочие, которыми должен обладать специалист для успешного выполнения своих функций.

Цель работы – предложить методику оценки профпригодности выпускников вуза, а также произвести настройку узлов для оценки подготовленности выпускников в выполнении конкретных трудовых функций профессиональных стандартов с решателем нейронная сеть, провести вычислительные эксперименты, сделать выводы об адекватности полученных результатов.

Представим задачу оценки профпригодности в виде иерархической модели (рис. 1), решение которой будет производиться на примере профессионального стандарта системного аналитика и направления подготовки бакалавров 09.03.03 «Прикладная информатика».

На первом этапе была проведена работа по установлению связи между дисциплинами учебного плана и трудовыми функциями профессионального стандарта [1]. Сопоставив полученные результаты с учебным планом направления подготовки бакалавров «Прикладная информатика», был составлен список дисциплин, которые могут готовить студентов к выполнению соответствующих трудовых функций.

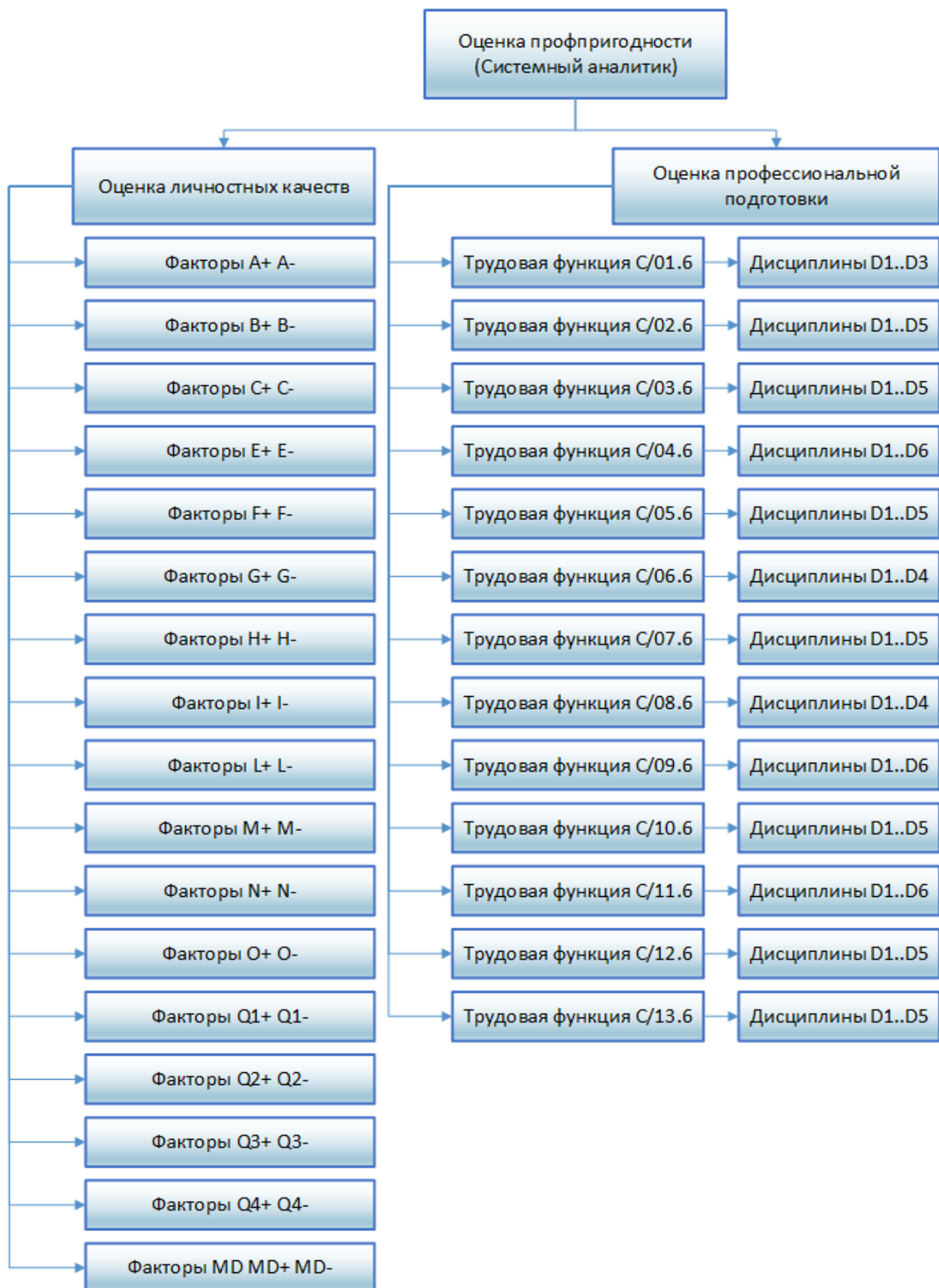


Рис.1. Иерархическая модель оценки профпригодности

Далее продолжилась работа с экспертами, в ходе которой для каждой трудовой функции был уточнен список дисциплин.

Рассмотрим детально задачу оценки подготовленности выпускника к выполнению трудовой функции «С/01.6 Планирование разработки или восстановления требований к системе» профессионального стандарта системного аналитика. В ходе проведения опроса экспертов было установлено, что к выполнению данной трудовой функции студентов готовят при изучении дисциплин Программная инженерия, Проектирование информационных систем, Проектный практикум.

Рейтинги студентов по данным дисциплинам и стали входами для нейронной сети. В качестве выходного параметра выступает оценка готовности студента к выполнению трудовой функции, которая была получена экспертным путем. Всего для обучающей выборки были собраны данные о выпускниках за три года (таблица 1).

Таблица 1

Фрагмент обучающей выборки

	X1 - Программная инженерия	X2 - Проектирование информационных систем	X3 - Проектный практикум	Y – Оценка готовности к выполнению трудовой функции
Студент 1	25	50	25	3
Студент 2	78	82	100	4
Студент 3	78	82	100	4
Студент 4	100	92	100	5
Студент 5	40	63	80	4
Студент 6	78	78	100	4
Студент 7	60	69	100	4
Студент 8	100	92	100	5

Таблица 2

Результаты обучения и тестирования нейронной сети

Метод обучения	Структ ура нейрон ной сети	Ошибка обучения			Ошибка обобщения		
		Максимальная	Минимальная	Средняя	Максимальная	Минимальная	Средняя
kPartan	3:3:1	1,282	0,001	0,282	1,917	0,012	0,497
kPartan	5:4:1	0,835	0,003	0,2	1,84	0,038	0,483
SCG	2:2:1	1,604	0,008	0,366	1,593	0,068	0,551
постраничный	5:4:1	1,149	0,009	0,325	1,941	0,041	0,414
постраничный	2:2:1	1,248	0,005	0,349	1,221	0,011	0,512
CGB	3:2:1	1,234	0,007	0,312	1,472	0,051	0,419
CGB	2:2:1	2,026	0,002	0,393	1,368	0,038	0,547
mPartan	5:4:1	1,151	0,001	0,232	1,643	0,022	0,504

mPartan	3:2:1	1,644	0,001	0,355	1,597	0,007	0,466
mPartan	4:4:1	0,852	0,001	0,277	1,807	0,068	0,53
BFGS	5:4:1	1,069	0,001	0,33	2	0,002	0,417
BFGS	2:2:1	1,817	0,018	0,433	1,562	0,068	0,624
BFGS	3:2:1	1,519	0,001	0,299	1,748	0,025	0,645
плавающий шаг	5:4:1	2,287	0,015	0,997	0,91	0,001	0,345
плавающий шаг	3:3:1	2,148	0,001	0,649	1,131	0,015	0,276
сопряженные градиенты	5:5:1	1,128	0,002	0,213	1,921	0,02	0,604
сопряженные градиенты	2:2:1	1,571	0,002	0,363	1,562	0,071	0,534
сопряженные градиенты	4:3:1	1,204	0,003	0,208	1,993	0,167	0,663

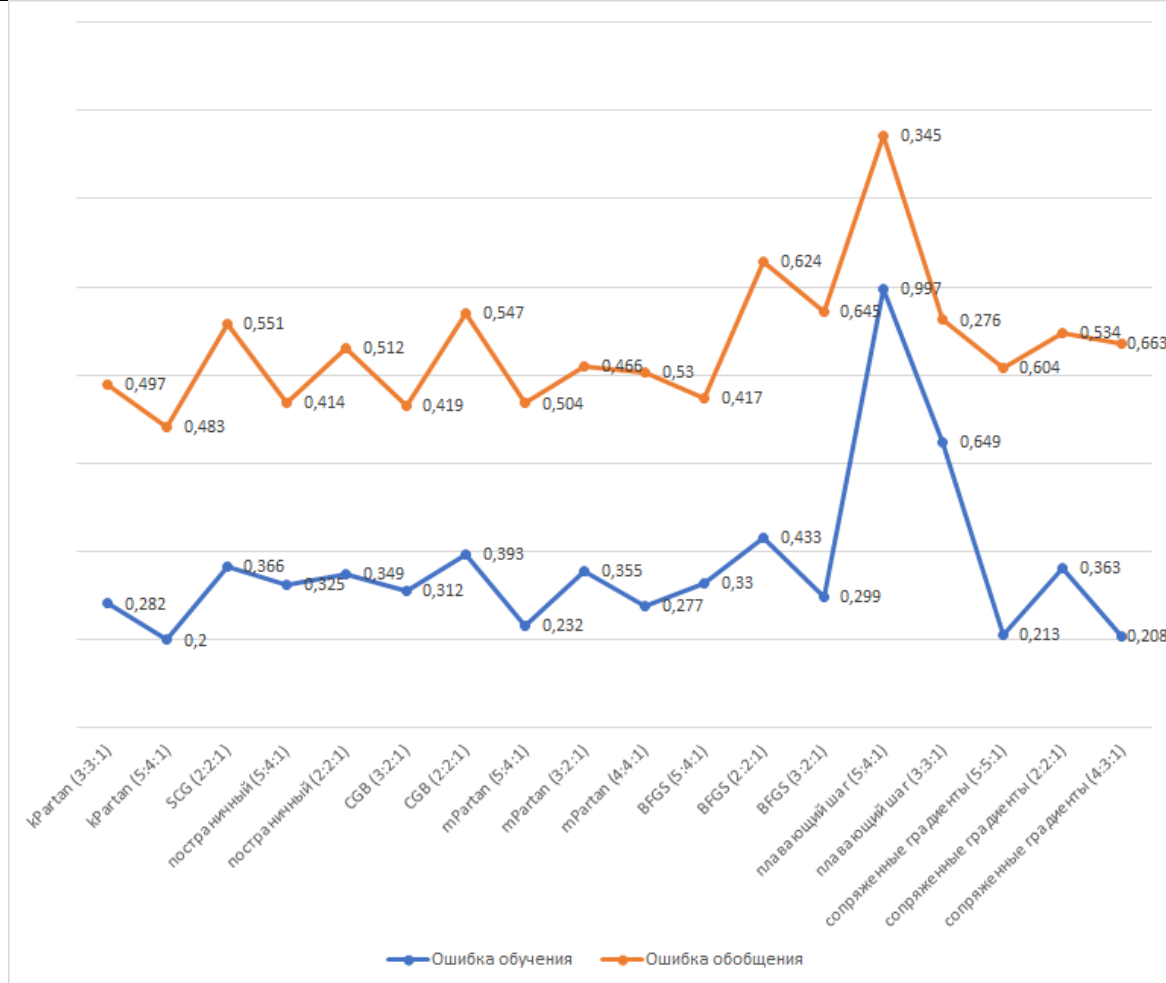


Рис.2. Графики изменения ошибок обучения и обобщения в зависимости от структуры сети и метода обучения

Перед обучением нейронной сети была проведена предобработка данных, которая включала анализ данных на наличие противоречивых примеров и пропусков.

Для выявления сети оптимальной структуры и метода обучения, наиболее эффективного для решения поставленной задачи были проведены эксперименты с сетями различной структуры с один скрытым слоем (таблица 2). Работы по обучению нейронной сети проводились в системе Нейро-Аналитик, разработанной кафедрой информационных систем в экономике АлтГТУ им. И.И. Ползунова.

Наилучшие результаты показал метод обучения «Плавающий шаг» со структурой сети 3:3:1, ошибка обобщения составила 0,276, а при структуре сети 5:4:1 – 0,345 (рис. 2). Если учесть что выходной параметр находился в диапазоне [0..5], то можно сделать выводы о погрешности полученных результатов не более 6%. Это позволяет сделать выводы об адекватности полученных результатов.

В дальнейшем планируется осуществить настройку решателей для остальных узлов модели, а также провести вычислительные эксперименты с разными нейроимитаторами для получения наиболее объективных результатов в решении данной задачи.

Список литературы

1. Пятковский О.И., Смыкова Н.В. Применение экспертных оценок для определения связей трудовых функций профессиональных стандартов и дисциплин учебного плана // Ползуновский альманах № 2/2016 Виртуальные и интеллектуальные системы (Выпуск посвящен Седьмой всероссийской научно-практической конференции ВИС-2016, 23 ноября). – Изд-во АлтГТУ. – 2012. – С.131-133.
2. «Трудовой кодекс Российской Федерации» от 30.12.2001 N 197-ФЗ (ред. от 01.07.2017). [Электронный ресурс] – URL: http://www.consultant.ru/document/cons_doc_LAW_34683/ - Загл. с экрана.
3. Федеральный закон «О внесении изменений в Трудовой кодекс Российской Федерации и статьи 11 и 73 Федерального закона "Об образовании в Российской Федерации» от 02.05.2015 N 122-ФЗ (последняя редакция). [Электронный ресурс] – URL: http://www.consultant.ru/document/cons_doc_LAW_178864/ - Загл. с экрана.

СИНТЕЗ МОДЕЛЕЙ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ МЕТОДОВ ВЫЧИСЛИТЕЛЬНОГО ИНТЕЛЛЕКТА И ФРАКТАЛЬНОГО АНАЛИЗА

Субботин С.А.

Запорожский национальный технический университет, г. Запорожье, Украина.

subbotin@zntu.edu.ua

Построение модели зависимости по ограниченному набору наблюдений с известной классификацией, является актуальной задачей во многих практических приложениях, где необходимо автоматизировать процесс принятия решений.

Пусть мы имеем исходную выборку $X = \langle x, y \rangle$ – набор S прецедентов о зависимости $y(x)$, $x = \{x^s\}$, $y = \{y^s\}$, $s = 1, 2, \dots, S$, характеризующихся набором N входных признаков $\{x_j\}$, $j = 1, 2, \dots, N$, и выходным признаком y . Каждый s -й прецедент представим как $\langle x^s, y^s \rangle$, $x^s = \{x_j^s\}$, где x_j^s – значение j -го входного признака x_j для s -го прецедента (экземпляра) выборки, y^s – значение выходного признака для s -го прецедента (экземпляра) выборки, $y^s \in \{1, 2, \dots, K\}$, $K > 1$, где K – число классов. Тогда задача синтеза модели зависимости $y(x)$ будет заключаться в определении такой структуры $F()$ и значений параметров ω модели, при которых будет удовлетворен критерий качества модели $f(F(), \omega, \langle x, y \rangle) \rightarrow opt$. Обычно критерий качества обучения распознающих моделей определяют как функцию ошибки модели E .

При построении моделей на основе методов вычислительного интеллекта [1] возникает необходимость предварительного сокращения размерности данных, что объясняется высокой итеративностью известных методов обучения моделей по прецедентам, а также необходимостью получить модель, обеспечивающую хорошее обобщение данных. При этом традиционно используемые методы отбора информативных признаков [1] и формирования выборок [1, 2] обладают одним общим недостатком – они не связаны непосредственно между собой и исходят с различных точек зрения на информативность признаков или экземпляров. С другой стороны, методы построения моделей [1] в своих целевых функциях – критериях качества обучения, как правило, не учитывают характеристики выборки. Это на практике может приводить к построению неоптимальных моделей.

Одним из перспективных направлений анализа данных является фрактальный анализ [3–10], ключевым понятием которого является фрактальная размерность – коэффициент, описывающий фрактальные структуры или множества на основе количественной оценки их сложности как коэффициент изменения в деталях с изменением масштаба.

Согласно [3] размерность Хаусдорфа-Бесиковича (Hausdorff – Besicovich dimension) определяется как $D \approx \log(n(l)) / \log(l^{-1})$, где $n(l)$ – минимальное число гиперкубов размером l , необходимых для покрытия образа. Одним из наиболее доступным способов определения размерности Хаусдорфа-Бесиковича является метод подсчета (box-counting method) [7], заключающийся в повторяющемся покрытии фрактального объекта гиперкубами равного размера и подсчетом каждый раз минимального числа гиперкубов, которые содержат точки образа. Последовательно уменьшая размер гиперкубов l , получают набор точек с координатами $(\log(n(l)), \log(l^{-1}))$, задающий кривую, наклон которой, определяемый с помощью линейной регрессии, является фрактальной размерностью:

$$D = \lim_{l \rightarrow 0} \frac{\log(n(l))}{\log(l^{-1})}.$$

Особенностью известных методов определения фрактальной размерности [3–10] является то, что размерность выборки и размерность модели, построенной на ее основе, определяются без связи друг с другом. Это ограничивает их практическое применение. Поэтому необходимо разработать методы оценивания фрактальной размерности, характеризующие свойства выборки и свойства обученной модели по выборке.

Целью данной работы являлось создание метода построения распознающих моделей на основе единого подхода – фрактального анализа, позволяющего как выборки данных, так и синтезируемые модели характеризовать с единой точки зрения.

Представим экземпляры обучающей выборки как точки в пространстве признаков. Тогда кластеры будут соответствовать компактным областям в пространстве признаков, которые будут объединяться в классы. Кластеры можно описывать различными геометрическими фигурами. Фрактальный анализ выборки в пространстве признаков можно осуществить, задав элементарную форму для выделения кластеров и варьируя размер кластера для разбиения выборки на фрагменты.

Для анализа фрактальной размерности выборки предлагается использовать следующий метод.

Этап инициализации. Задать обучающую выборку $\langle x, y \rangle$ и L – число интервалов, на которые разбиваются диапазоны значений признаков.

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал $[0, 1]$: $x_j^s = (x_j^s - x_j^{\min}) / (x_j^{\max} - x_j^{\min})$, где $x_j^{\max}, x_j^{\min}$ – соответственно, максимальное и минимальное значения признака x_j ;

Этап кластеризации. Разбить диапазон значений каждого признака на L интервалов длиной l : $l = L^{-1}$. Сформировать кластеры как прямоугольные блоки на пересечении интервалов разных признаков.

Этап анализа данных. Определить число экземпляров, попавших в каждый прямоугольный гиперблок, образованный интервалами признаков $n_{i,q}$

Определить число экземпляров k -го класса, попавших в каждый прямоугольный гиперблок, образованный интервалами признаков $n_{i,q,k}$

Определить число гиперблоков со стороной размером l , покрывающих k -й класс выборки в пространстве N признаков: $n_{(k)} = \sum_{i=1}^N \sum_{q=1}^L \{1 \mid n_{i,q,k} > 0\}$.

Определить число гиперблоков со стороной размером l , покрывающих выборку в пространстве N признаков:

$$n(l) = \sum_{i=1}^N \sum_{q=1}^L \{1 \mid n_{i,q} > 0\} = \sum_{i=1}^N \sum_{q=1}^L \left\{ 1 \mid \sum_{k=1}^K n_{i,q,k} > 0 \right\}$$

Этап определения фрактальной размерности. Определить при заданном l фрактальную размерность k -го класса, $k = 1, 2, \dots, K$: $D^{(k)} = \log(n_{(k)}) / \log(l^{-1})$.

Определить фрактальную размерность выборки при заданном l : $D = \log(n(l)) / \log(l^{-1})$.

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ими пространство признаков. Единственным управляемым параметром метода является задаваемое число интервалов L , на которые разбиваются диапазоны значений признаков. Очевидно, что число кластеров $Q \geq K$, а $Q = L^N$, причем для каждого признака $L \geq 2$. Для обеспечения обобщающих свойств кластеров введем ограничение $Q \leq NS$. Таким образом, получим: $2 \leq L \leq \sqrt[N]{NS}$. Заметим, что минимальным шагом для варьирования значений L является 1. Если значение верхнего предела $\sqrt[N]{NS}$ оказывается

меньше 2, то его можно заменить на S . Это связано с тем, что на оси каждого признака будет не более чем S точек и разбиение их на более чем S интервалов, очевидно, будет приводить к возникновению пустых интервалов. При больших значениях N задание числа разбиений порядка S по каждому признаку приведет к выделению огромного числа блоков порядка S^N , что сделает вычисления крайне трудоемкими, а в ряде случаев и практически не реализуемыми. Поэтому рационально в этом случае ограничить значение верхнего предела значений L величиной порядка $\text{round}(\lg(S))$, где round – функция округления к ближайшему целому.

Определение показателя D при малых значениях L потребует больших затрат вычислительных ресурсов и ресурсов памяти ЭВМ, чем при больших значениях L . Однако точность анализа для малых значений L будет ниже, в то время как уровень обобщения будет выше, чем для больших значений L .

Достоинством предложенного метода и определяемого на его основе показателя качества выборки является то, что они не зависят от метода синтеза модели и результатов его работы и позволяют оценивать свойства отдельно взятой выборки. Недостатками предложенного метода являются неопределенность в выборе значения параметра L , а также отсутствие связи метода с качеством синтезируемой модели.

Для устранения отмеченных недостатков предлагается использовать метод определения фрактальной размерности входного множества для обучения распознающей модели.

Этап инициализации. Задать обучающую выборку $\langle x, y \rangle$, метод синтеза модели, критерий качества обучения модели как функцию ошибки E , а также максимально приемлемое значение ошибки ε .

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал $[0, 1]$.

Этап формирования и анализа разбиения данных. Последовательно изменяя значение $L = 2, \dots, S$: определить длину интервала l ; квантовать признаки выборки, разбив диапазоны их значений на L интервалов; определить число гиперблоков со стороной размером l , покрывающих выборку в пространстве N признаков, $n(l)$; построить распознающую модель net с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε ; оценить ошибку E построенной распознающей модели net .

Этап определения фрактальной размерности. Для всех l , для которых ошибка модели является приемлемой, определить фрактальную размерность данных относительно точности (ошибки) синтезированной модели:

$$D_{net} = \left\{ \frac{\log(n(l))}{\log(l^{-1})} \middle| E(net) \leq \varepsilon \right\}.$$

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ими пространство признаков. Единственным управляемым параметром метода является задаваемое граничное значение ошибки ε модели. Очевидно, что чем меньше будет заданное ε , тем более детальной должна быть модель, то есть потребуются выделить большее число кластеров Q , а, следовательно, и большим будет число L . Соответственно, с уменьшением заданного ε будут возрастать затраты вычислительных ресурсов и ресурсов памяти ЭВМ на анализ выборки.

Достоинством предложенного метода и определяемого на его основе показателя качества выборки является то, что они связаны с показателем качества синтезируемой модели, а также в автоматическом режиме устанавливает оптимальное значение параметра L . Недостатками предложенного метода является неопределенность в выборе значения параметра ε , а также его зависимость от качества обучения и принципов

функционирования модели, по которой он определяется. Также следует отметить, что функция ошибки, используемая в методе, является лишь одной из характеристик синтезируемой модели, но не учитывает её размерность и обобщающие свойства.

Поэтому фрактальную размерность обученной модели предлагается определять на основе изложенного ниже метода, учитывающего размерность модели.

Размерность модели – число величин, представляющих в модели параметры и характеристики. Для распознающей модели оценим размерность числом ее параметров w и числом структурных элементов e . Для разных типов моделей эти величины могут существенно отличаться. Однако, как правило, w зависит от e , причем $w \gg e$. Поэтому предлагается для оценки размерности распознающей модели ограничиться числом её параметров w . Определим максимальное допустимое с точки зрения обобщения значение w как размерность входа обучающего множества $w_{\max} = NS$, а минимально допустимое значение $w_{\min} > 0$.

Этап инициализации. Задать обучающую выборку $\langle x, y \rangle$, метод синтеза модели, критерий качества обучения модели как функцию ошибки E , а также максимально приемлемое значение ошибки ε . Задать $w = w_{\max}$.

Этап нормирования выборки. Если значения признаков ненормированы, то их следует пронормировать, отобразив на интервал $[0, 1]$.

Этап формирования и анализа разбиения данных. Последовательно изменяя значение $L = 2, \dots, S$: определить длину интервала l ; квантовать признаки выборки, разбив диапазоны их значений на L интервалов; определить число гиперблоков со стороной размером l , покрывающих выборку в пространстве N признаков, $n(l)$; построить распознающую модель $net(w)$ с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε ; оценить ошибку E построенной распознающей модели net .

Этап минимизации сложности модели. Найти минимальное l при котором точность построенной модели $net(w)$ является приемлемой. Последовательно изменяя значения $w = w_{\min}, \dots, w_{\max}$ построить модель $net(w)$ с помощью заданного метода, минимизируя функцию ошибки E до достижения приемлемого уровня ε . Для всех w , для которых ошибка модели является приемлемой, определить фрактальную размерность данных относительно точности (ошибки) синтезированной модели при текущем w :

$$D_{net(w)} = \left\{ \frac{\log(n(l))}{\log(l^{-1})} \middle| E(net(w)) \leq \varepsilon \right\}.$$

Для $D_{net(w)} > 0$ определить такое минимальное w , при котором при минимальном L ошибка модели E будет приемлемой:

$$D_{net^*} = \left\{ \frac{D_{net(w)}}{\log(w^{-1})} \middle| E(net(w)) \leq \varepsilon \right\} = \left\{ \frac{\log(n(l))}{\log(l^{-1}) \log(w^{-1})} \middle| E(net(w)) \leq \varepsilon \right\}$$

Данный метод оперирует прямоугольными блоками одинакового размера, покрывая ними пространство признаков. Управляемыми параметрами метода является задаваемое граничное значение ошибки ε модели, а также число параметров модели w . Очевидно, что чем меньше будет значение w , тем выше может быть уровень обобщения модели, однако и тем будет сложнее построить модель с требуемой точностью ε .

Для исследования комплекса предложенных показателей выборок и моделей, основанных на фрактальной размерности, они были программно реализованы. Разработанное программное обеспечение использовалось для проведения вычислительных экспериментов по исследованию применимости предложенных показателей на примере решения задачи автоматической классификации ирисов Фишера по признакам [11], выборка данных для которой визуализирована на рис. 1.

В результате проведенных экспериментов для задачи классификации ирисов Фишера оценка фрактальной размерности выборки составила $D = 0,59034$, а оценки фрактальных размерностей классов: $D^{(1)} = 0,68223$, $D^{(2)} = 0.6212$, $D^{(3)} = 0.53407$. Графики зависимостей от l^{-1} в логарифмической системе координат для всей выборки и классов приведены на рис. 1а и 1б, соответственно. На рис. 1б разные классы кодируются маркерами разных размеров. На рис. 1в и 1г приведены схематические графики зависимостей $n(l)$ выборки от l^{-1} в логарифмической системе координат для разных заданных значений ε и получаемых значений E (рис. 1б), а также размерностей формируемой выборки S' и N' (рис. 1г).

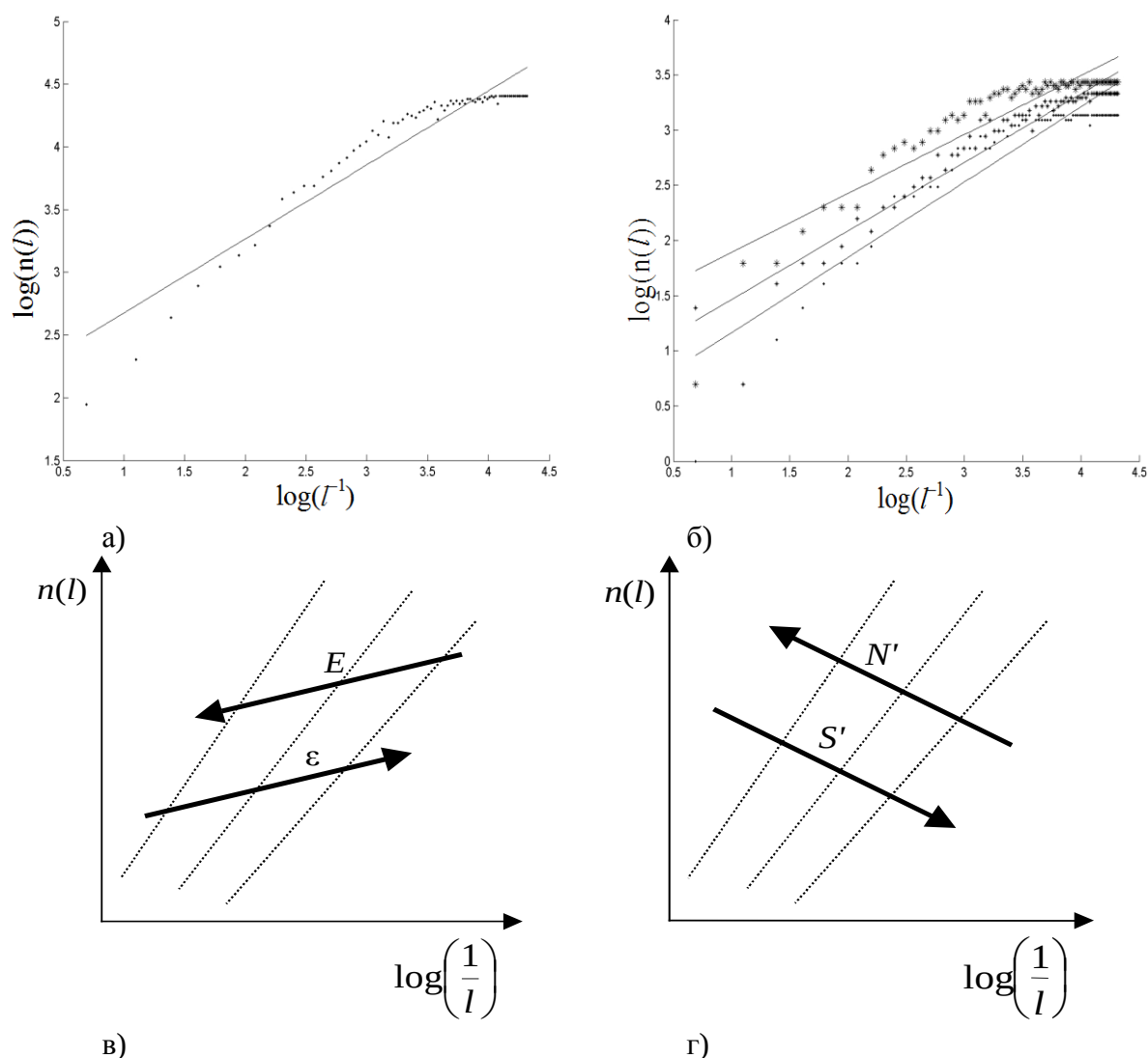


Рис.1. Графики зависимостей от l^{-1} в логарифмической системе координат: а) – $n(l)$ выборки, б) – $n(l)$ классов; в) – схематические для $n(l)$ выборки, г) – схематические для $n(l)$ классов

Проведенные эксперименты подтвердили работоспособность предложенного математического обеспечения и позволяют рекомендовать его для использования на практике при решении задач построения моделей по прецедентам.

Как видно из рис. 2 и рис. 4, а также табл. 1, предложенные показатели фрактальной размерности позволяют хорошо показать различия между классами. Эти показатели можно использовать в методах формирования выборок, определяя критерии

качества формируемых подвыборок на основе предложенных показателей фрактальной размерности. Если формируемая подвыборка или ее классы по показателям фрактальной размерности существенно отличаются от аналогичных показателей исходной выборки, то возможно, что полученная выборка не обладает репрезентативностью относительно исходной выборки. Также при сравнении нескольких подвыборок-кандидатов предложенные показатели могут использоваться как меры их качества: из подвыборок-кандидатов следует отдавать предпочтение той, которая будет иметь показатели фрактальной размерности, наиболее близкие по значениям к показателям исходной выборки.

Как видно из рис. 1, изменение заданных составляющих размерности формируемой подвыборки (числа признаков N' и числа экземпляров S'), а также E и ε влияют на положение прямой, соединяющей точки зависимости $n(l)$.

Перспективы дальнейших исследований могут заключаться в создании параллельных методов расчета комплекса предложенных показателей, оптимизации их программных реализаций, а также экспериментальном исследовании предложенных показателей на большем комплексе практических задач разной природы и размерности.

Список литературы

1. Jensen R. Computational intelligence and feature selection: rough and fuzzy approaches / R. Jensen, Q. Shen. – Hoboken: John Wiley & Sons, 2008. – 339 p.
2. Chaudhuri A. Survey sampling theory and methods / A. Chaudhuri, H. Stenger. – New York: Chapman & Hall, 2005. – 416 p.
3. Roberts A. Unbiased estimation of multi-fractal dimensions of finite data sets / A. Roberts, A. Cronin // *Physica A: Statistical Mechanics and its Applications*. – 1996. – Vol. 233, № 3-4. — P. 867-878.
4. Evaluating the fractal dimension of profiles / [B. Dubuc, J. Quiniou, C. Roques-Carnes, C. Tricot, S. Zucker] // *Physical Review*. – 1989. – T. 39, № 3. – P. 1500–1512.
5. Cheng Q. Multifractal Modeling and Lacunarity Analysis / Q. Cheng // *Mathematical Geology*. – 1997. – Vol. 29, № 7. – P. 919–932.
6. Eftekhari A. Fractal Dimension of Electrochemical Reactions / A. Eftekhari // *Journal of the Electrochemical Society*. – 2004. – Vol. 151, № 9. – P. E291–E296.
7. Li J. An improved box-counting method for image fractal dimension estimation / J. Li, Q. Du, C. Sun // *Pattern Recognition*. – 2009. – Vol. 42, № 11. – P. 2460–2469.
8. Crişan D. A. Fractal dimension spectrum as an indicator for training neural networks / D. A. Crişan, R. Dobrescu // *Universitatea Politehnica Bucuresti Sci. Bull. Series C*. – 2007. – Vol. 69, № 1. – P. 23–32.
9. Takens F. On the numerical determination of the dimension of an attractor / F. Takens // *Dynamical Systems and Bifurcations : Workshop, Groningen, 16–20 April 1984 : proceedings* / [eds.: Braaksma B., Broer H. W., Takens F.]. – Berlin: Springer, 1985. – P. 99–106. – (Lecture Notes in Mathematics, Vol. 1125).
10. Zong-Chang Y. Establishing structure for artificial neural networks based-on fractal / Y. Zong-Chang // *Journal of Theoretical and Applied Information Technology*. – 2013. – Vol. 49, № 1. – P. 342–347.
11. Fisher Iris dataset [Electronic resource]. – Access mode: <https://archive.ics.uci.edu/ml/datasets/Iris>

ПРОВЕРКА ГИПОТЕЗЫ ОБ УРОВНЕ РАЗВИТИЯ КОМПЕТЕНТНОСТИ С ИСПОЛЬЗОВАНИЕМ КРИТЕРИЯ ШОРТЛИФФА И БЬЮКЕНЕНА

В.А.Углев

Сибирский федеральный университет, uglev-v@yandex.ru

Анализ результатов решения контрольно-измерительных материалов при обучении должен завершаться выставлением оценки/отметки и формированием рекомендаций. Эти результаты обычно представляют собой вектор ответов учащегося на отдельные задания (статистический ряд) или траекторию их выполнения (временной ряд). По сути, в бланках ответов/наблюдений или протоколах работы автоматизированных обучающих систем хранятся вектора с цифрами, кодирующими действия ученика [1]. Компетенции, которые в современном подходе к обучению ложатся в основу учебных программ, так же являются объектом оценки и предполагают создание соответствующих методов. Не смотря на то, что компетенция является латентной переменной [2], её измерение классическими методами параметрической статистики не корректно, а традиционные методы непараметрической статистики, применяемые в педагогике [3], нуждаются в адаптации к специфике предмета измерения. В работах [4, 5] был кратко описан подход к оценке уровня развития компетентностей (УРК), опирающийся на проверку гипотез с помощью критерия Шортлиффа (Shortliffe E.) и Бьюкенена (Buchanan B.). Рассмотрим этот критерий подробнее с учётом специфики процесса проверки гипотез об УРК в автоматизированном режиме.

Проверка гипотез по результатам обработки вектора ответов учащегося на задания для оценки уровня остаточных знаний может быть осуществлена простым подсчетом баллов за каждый ответ. Но латентные переменные, такие как компетентности, таким образом оценить проблематично. Тогда можно прибегнуть к методам, базирующимся на экспертных оценках. К ним, например, можно отнести подходы Демпстера-Шефера, Криса Нейдлора, а так же Шортлиффа и Бьюкенена [6]. Последний критерий имеет свою специфику и нашел широкое применение в экспертных системах. Так, достоинствами данного критерия являются возможность учитывать противоречия (каждое событие ряда может одновременно свидетельствовать как в пользу одних, так и не в пользу других гипотез), а так же одновременно проверять ряд гипотез. Среди недостатков критерия Шортлиффа и Бьюкенена можно отметить необходимость составления матрицы адекватных экспертных оценок.

Процесс работы с УРК можно условно разделить на три больших этапа: подготовительный, измерительный, оценочный (собственно проверка гипотез) и завершающий. Рассмотрим каждый из них, выделив шаги, касающиеся рассматриваемого подхода. Начнем с *подготовительного*.

Первым шагом при организации работы с УРК является формализация компетенций из учебного плана. Для этого осуществляется следующая последовательность действий: выделение *структуры компетенции* и *этапов* её освоения. Под структурой компетенции понимается деление её на более менее независимые фрагменты знаний и отмечается глубина их проработки. Например, имеется компетенция «способность к абстрактному мышлению, анализу, синтезу», которую можно разбить на такие составляющие, как «способность анализировать», «способность синтезировать» и «способность абстрагироваться». Далее, в каждой составляющей выделяется фрагмент, отвечающий специфике обучения. Например, покажем для первой составляющей из приведенного выше примера фрагменты и их характеристики относительно учебного плана (см. таблицу 1). Согласно действующим подходам к декомпозиции учебного материала, каждый фрагмент будет иметь глубину освоения *знать/уметь/владеть*. Для выделения этапов осуществляется анализ учебного

плана и отмечается, на каких дисциплинах, в каких семестрах и на каких темах задействованы те или иные фрагменты компетенций с учётом преемственности глубины освоения (от вводного *знать* до завершающего *владеть*). Таким образом, в любой момент времени в процессе обучения можно говорить о фиксации текущего этапа формирования компетенции и «зоны ответственности» отдельных дисциплин в её освоении.

Таблица 1. Формализация фрагмента из структуры компетенции

№	Фрагмент	Уровень	Дисциплины (семестры)	Темы в составе дисциплин
1	осуществлять декомпозицию систем	Знать	1. Системный анализ и исследование операций (2,3) 2. Глобальная и многокритериальная оптимизация (1, 2) 3. Основы теории активных систем (2)	1.1-1.12, 2.2, 2.4-2.6, 2.9, 3.2, 3.3, 4.3
		Уметь		1.5-1.10, 2.3, 2.5, 2.8, 3.3, 3.4
		Владеть		1.8, 1.10, 2.5-2.7, 2.10, 3.5
2	представлять систему в зависимости от целевого рассмотрения	Знать	4. Методы искусственного интеллекта (3)	1.3-1.6, 2.4, 3.2, 3.4, 4.3
		Уметь		1.5, 1.6, 3.2, 4.3
		Владеть		1.6, 3.2
3	выделять структуру, функции и роли	Знать		1.2-1.12, 3.2-3.4
		Уметь		1.4-1.7, 3.4
		Владеть		1.5, 1.7, 3.4
4	владеть логическими и математическими методами анализа	Знать		1.6, 1.7, 1.9-1.12, 2.1-2.9, 3.5, 4.3, 4.6-4.8
		Уметь		1.7, 1.9-1.11, 2.5, 2.6, 2.9, 3.5, 4.6, 4.7
		Владеть		1.7, 1.10, 2.6, 2.9, 3.6, 4.6

Вторым шагом является составление контрольно-измерительных материалов (КИМов). Для этого следует соотнести для каждой темы, подлежащей контролю, глубину развиваемой компетенции и соответствующей ей формы заданий: от тестовых заданий закрытой формы (для проверки знания понятийного аппарата), до практических расчетных задний для выявления корректности и оптимальности действий ученика (при демонстрации овладения тем или иным фрагментом компетенции). Каждый вариант ответов заданий и тема/дидактическая единица будут формировать двудольный граф (подробнее см. [1]), по которому в последствии можно сформировать множество актуальных для контроля КИМов.

Третьим шагом будет составление матрицы экспертных оценок, позволяющих рассчитать коэффициенты уверенности в УРК или её структурной части по Шортлиффу и Бьюкенену [7]. Для этого формируется матрица смежности так, как показано в [5]. Значения экспертных оценок для каждого задания составляет учитель-предметник или их группа (в этом случае применяются приёмы согласования экспертных мнений, например, метод нечёткой логики). Специфика составления заданий подробно описана в [5] и [8], а особенности формирования оценок кратко представлены в таблице 2.

Таблица 2. Специфика формирования экспертных оценок

Тип задания	Форма ответа	Оценка эксперта	Степень автоматизации
Тестовое задание закрытой формы или расчётная задача	Выбор одного или комбинации ответов	По схожести с эталоном	Высокая
Тестовое задание открытой формы, эссе или устный ответ	Ввод значения по шаблону или в произвольной форме	По корректности содержания ответа	Низкая
Выполнение задания в виртуальной лаборатории	Формирование графа выполнения действий и фиксация параметров учебной модели	По корректности действий с учётом этапа решения	Средняя
Выполнение практической работы (например, в очках дополненной реальности)			
Задание проективного типа	Выбор одного ответа	По значимости для компетенций	Высокая

На этапе *измерения* и фиксации осуществляется выполнение следующих шагов: формирование выборки заданий, соответствующая текущему моменту обучения по сложности и содержанию; предъявление заданий ученику и фиксация ответов. При необходимости реализовать индивидуальную учебную траекторию, как состав, так и форма предъявления заданий могут быть адаптированы под конкретного учащегося (включая автоматизированный режим [9]). Фиксация ответов учащегося производится автоматически, но при решении отдельных видов заданий (оценка эссе, ответов в открытой форме и пр.) их может вводит учитель после непосредственной проверки.

Этап *оценки*, являющийся ключевым, будет состоять из следующих пяти шагов, которые представлены далее.

Первым шагом является подсчет баллов за решения всех КИМов, оценивающий знаниевую составляющую по любой из методик, не имеющих отношение к задаче оценки УРК. Обычно, за одно задание выставляется балл из интервала от 0 (совершенно не верно/нет ответа) до 1 (полностью верный ответ). Как показано в [5], отдельные типы заданий могут не содержать баллов правильности и, как следствие, не включаться в процесс оценки уровня остаточных знаний.

Вторым шагом является расчет накопленных экспертных оценок, участвующих в проверке гипотезы о развитии той или иной компетентности у ученика. Они так же выставляются в пределах от 0 до 1 или от -1 до 1 (настраиваемый параметр), но отражают не правильность решения, а степень проявления учащимся той или иной компетентности.

На *третьем* шаге производится, собственно, расчет значения коэффициента уверенности C_f состоит из двух компонент [7]: оценка меры доверия, подтверждающей

состояние (MB) гипотезы, и меры не доверия (MD), свидетельствующей в пользу её опровержения по (1).

$$MB[H, x] = \frac{p(H/x) - p(H)}{1 - p(H)} \quad \text{и} \quad MD[H, x] = \frac{p(H) - p(H/x)}{p(H)}. \quad (1)$$

Комбинация MB и MD по (2) даёт итоговую оценку гипотезы об УРК (подробнее технология расчёта была приведена в [1] и [5]).

$$Cf(a_k, q_a \wedge q_{\bar{a}}) = MB(a_k, q_a) - MD(a_k, q_{\bar{a}}). \quad (2)$$

На *четвёртом* шаге производится нормирование (масштабирование) значений оценок каждого УРК, с целью сравнимости результатов различных компетентностей между собой и возможности корректного отслеживания динамики их значений.

Пятый этап предполагает фиксацию результатов и, при необходимости, перевод количественных значений УРК в отметки, с выделением качественной характеристики уровня их развития.

Завершающий этап состоит из предъявления оценки/отметки ученику (обычно в разрезе знаний, а не компетенций); комментирование результатов, формирование рекомендаций к дальнейшему обучению. По отношению к учителю это дополнительно функции визуализации (моментальный срез или динамика), обобщения по группе, расчета дополнительных статистических показателей и развернутое пояснение того, почему были автоматически предложены те или иные рекомендации [10]. Все эти функции поддаются комплексной автоматизации и базируются на технологии экспертных систем.

Результаты практического применения представленного подхода были использованы в учебном процессе при подготовке ряда специальностей (бакалавры, магистры) в СФУ и КГПУ. Для примера, на рисунке показан результат оценки УРК трёх студентов по дисциплине «Имитационное моделирование» (каждая ось соответствует одной из компетенций учебного плана, 0 – граница неопределённости). Данный подход так же был применен при осуществлении отбора абитуриентов и при проведении аттестационных мероприятий.

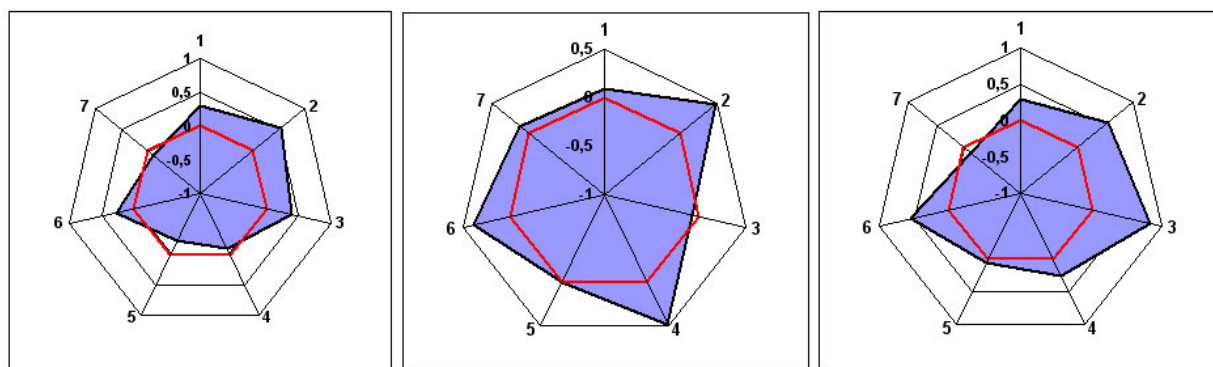


Рис. Результаты оценки УРК трёх студентов

В заключение следует отметить, что качественный анализ и оценивание уровня развития компетентностей базируется на качественной подготовке контрольно-измерительных материалов, тесно интегрированных в учебный план в рамках

компетентностный модели, а так же на корректности обработки статистики ответов. Применение в процессе проверки гипотез, на базе коэффициента уверенности Шортлиффа и Бьюкенена, позволяет повысить управляемость учебный процессом, а так же аргументированность принимаемых учителем решений. Более того, многие решения могут быть выработаны автоматически.

Список литературы

1. Углев В.А. Методика извлечения знаний о компетентностях из протоколов автоматизированной обучающей системы подготовки операторов // Нейроинформатика, ее приложения и анализ данных: XXI Всероссийский семинар.–Красноярск,2013.–С.144–150.
2. Лазарсфельд П.Ф. Логические и математические основания латентноструктурного анализа // Математические методы в современной буржуазной социологии. – М.: Прогресс, 1966. – С. 344–401.
3. Грабарь М.И., Краснянская К.А. Применение математической статистики в педагогических исследованиях. Непараметрические методы. – М.: Педагогика, 1977. – 136 с.
4. Uglev V.A., Ustinov V.A. The new competencies development level expertise method within Intelligent Automated Educational Systems // Trends in Practical Applications of Heterogeneous Multi-Agent Systems. The PAAMS Collection. Advances in Intelligent Systems and Computing, - 2014. – Vol. 293. – pp. 157-164 (DOI 10.1007/978-3-319-07476-4_19).
5. Углев В.А., Добронец Б.С. Методика автоматизированного измерения и оценки уровня развития компетентностей // Информатика и образование.– № 2. – 2017. – С. 61-65.
6. Джексон П. Введение в экспертные системы. – М.: Вильямс, 2001. – 624 с.
7. Buchanan B., Shortliffe E. Rule-based Expert Systems. – New York: Addison-Wesley, 1984. – 748 p.
8. Углев В. А. Экспериментальная отработка методики автоматизированной оценки уровня развития компетентностей // Информатизация образования и методика электронного обучения: материалы I Международной научной конференции. – Красноярск: СФУ, 2016. – С. 107-112.
9. Углев В.А. Возможности индивидуализации обучения в автоматизированных обучающих системах на базе когнитивных карт диагностики знаний // Перспективные методы и средства интеллектуальных систем: Материалы всероссийского научно-практического семинара и школы молодых ученых. – Новосибирск: Изд-во НГТУ, 2015. – С. 41-45.
10. Углев В.А. Выявление закономерностей и их усиление при обработке протоколов работы автоматизированных обучающих систем с помощью синтеза когнитивных карт диагностики знаний // Проблемы управления в социальных системах. – Том 6, вып. 9. – 2013. - С. 108-121.

СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПО ОТБОРУ АБИТУРИЕНТОВ В МАГИСТРАТУРУ С ЭЛЕМЕНТАМИ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

В.А. Углев, К.Г. Шаталов, С.В. Холодилов, В.А. Холодилова

Сибирский федеральный университет, uglev-v@yandex.ru

Качество организации и проведения процедуры отбор абитуриентов в высшие учебные заведения во многом определяет результат обучения. Для уровня магистратуры это оправдано в большей степени, т.к., с одной стороны, это массовая форма получения образования (в отличие от аспирантуры), а, с другой стороны, ориентирована на проявление и развитие компетентностей, связанных с научной работой. Как правило, зачисление осуществляется по результатам собеседования, с учётом некоторой априорной информации о достижениях абитуриента на этапе поступления. Но в случаях, когда конкурс очень мал (до ситуаций недобора при непрофильном контингенте абитуриентов⁶), а процент отчисления в процессе обучения велик (до 80%), опираться исключительно на априорную информацию не достаточно. Иногда важно не столько то, где и как обучался абитуриент, а то, насколько он мотивирован обучаться, с учётом специфики организации обучения по конкретной специальности. Таким образом, было принято решение исследовать возможности построение модели абитуриента, по которой можно было бы прогнозировать его потенциальную способность успешно завершить обучение. Полученную методику отбора и средства автоматизации мы ориентировали на работу приёмной комиссии в магистратуре межинститутской базовой кафедры «Прикладная физика и космические технологии» Сибирского федерального университета (г. Железногорск), ориентированной на предприятие «Информационные космические системы» им. Акад. М.Ф. Решетнёва».

Собственно проблемой отбора абитуриентов занимались многие: нами были проанализированы работы Е.К. Гитман, В.С. Сенашенко, А.В. Гаврилова, О.А. Грачиной, Т.В. Васильевой, И.С. Хлопова, Е.М. Родионова, Л.В. Ведерниковой, С.А. Бобинкой, С.В. Тархова, Е.В. Мироновой, А.В. Ивашиной и др. Практически все указанные подходы проводятся в *три этапа* (анкетирование, входной тест/решение кейсов или задач, собеседование) и многие сопровождаются средствами автоматизации. Они ориентированы на различные уровни образования или трудовой деятельности и не выполняют прогностическую функцию, т.е. оценивают исключительно априорные данные. Попытаемся проанализировать имеющиеся данные о работе приёмной комиссии, выявив некоторые закономерности обучения магистрантов, и предложить методику, которую и применим при отборе абитуриентов. Рассмотрим подробнее все этапы отбора, выбрав в качестве объекта исследования специальность 09.04.01.03 (программа обучения «Информационные системы космических аппаратов и центров управления полётами»), введя *четвёртый этап* - прогнозирование.

⁶ Как показывает наш опыт, для отдельных специальностей на поступление могут приходить абитуриенты, профиль базового образования которых (специалитет или бакалавриат) не является профильным (до 60% от потока абитуриентов).

Анализ статистических данных за период с 2012 по 2017 годы включал 72 прецедента, из которых 55 были зачислены, а 15 успешно окончили обучение⁷. На основании анкетных данных, результатов входного тестирования и статистики обучения была сформирована *параметрическая модель абитуриента*, включающая 63 параметра [1]. Она делилась на две части: вектор априорных данных, получаемых в процессе поступления (26 показателей по анкете, результатам оценки уровня знаний и уровня развития компетентностей, сводных оценок экспертной комиссии), и вектора апостериорных данных, фиксируемых в процессе обучения (37 показателей из зачетной книжки, специфики темы научной работы и руководителя, активности студента, результатов сдачи им госэкзамена и защиты ВКР). Очевидно, что вектор апостериорных показателей имелся не у всех (отчислены в первом семестре), тем более в полном объеме (отчислены позднее, см. рис. 1).

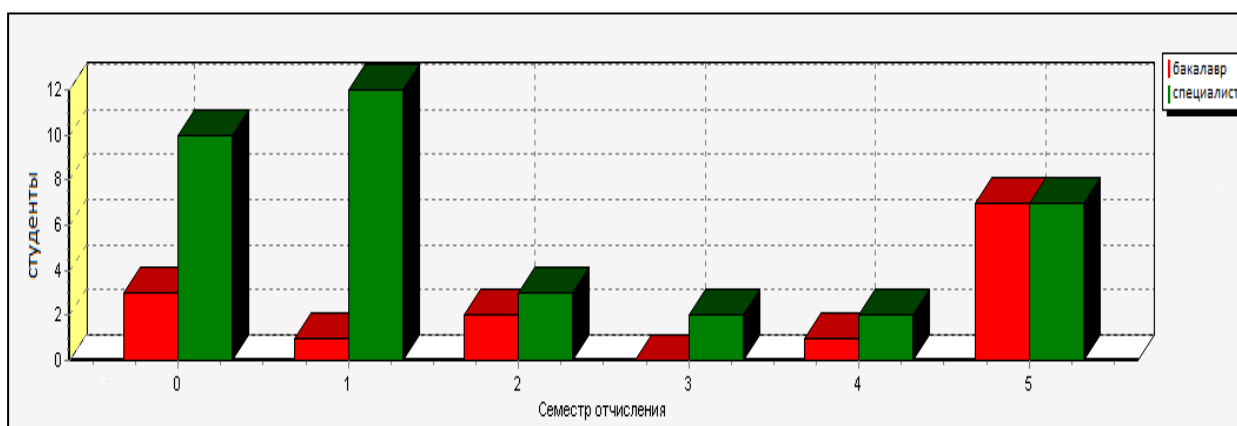


Рис 1. Статистика отчислений по семестрам⁸

Далее был осуществлен анализ данных разными методами и по каждому методу были выявлены те или иные *закономерности*:

- искусственная нейронная сеть с учителем (программа NeuroPro) – показано, что компетенции, отвечающие за самостоятельность и общую ИТ-грамотность, имеют для успешной защиты диссертации наибольшую значимость, а компетенции, связанные с кругозором и специализации в ракетно-космической отрасли – наименьшую;

- корреляционный анализ (программа MS Excel) – показано, что наличие профильного базового образования и опыта студенческой «научной» работы скорее характеризую студентов, не склонных серьезно относиться к выполнению диссертационной работы; за то уровень мотивации и выбор руководителя в значительной степени определяют успех проведения научного исследования;

- анализ классов (программа ViDaExpert [2]) – показал, что данные имеют достаточно чёткие кластеры успешно завершивших обучение (красные) и не успешно (синие), при этом данные кластеры не являются линейно сепарабельными (см. рис. 2);

⁷ Следует отметить, что на текущий момент ещё 15 человек находятся в процессе обучения и из них далеко не все выйдут на защиту магистерской диссертации, т.к. на данной специальности (академическая магистратура) достаточно требовательно относятся к качеству подготовки.

⁸ Здесь отчисление в нулевом семестре – «отчисление» не прошедших отбор абитуриентов; в пятом семестре – отчисление в связи с успешным окончанием обучения.

- многомерный анализ данных (компоненты кубов данных Decision Cube в Borland C Builder) – подтвердил ранее выявленные закономерности и позволил сделать вывод о том, что наибольший риск отчисления у специалистов – первый семестр, а у бакалавров – второй (см. рис. 1).

На основании этих и прочих закономерностей, характерных для данной специальности в г. Железногорске была предложена методика отбора, позволяющая не только оценить, но и спрогнозировать результаты обучения (сама методика изложена в [3]). Последовательность приёма стала следующей:

- на первом этапе производилось анкетирование, фиксирующее предпочтения абитуриента и некоторые персональные данные (включая отдельные психологические характеристики);

- на втором этапе с помощью специально подготовленных контрольно-измерительных материалов измерялся уровень остаточных знаний, уровень развития компетентностей (см., например, рис. 3), уровень социальной желательности, а так же соответствие замеров самооценке абитуриента;

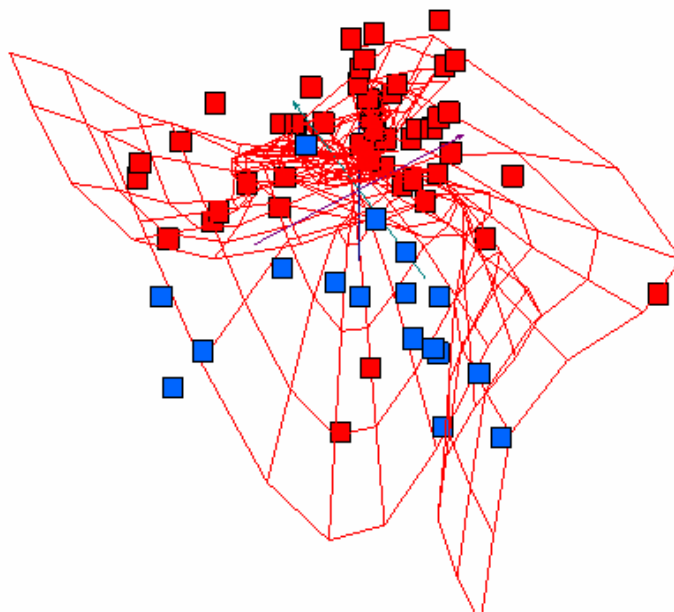


Рис 2. Упругая карта результатов обучения

- на третьем этапе производилось очное собеседование, на котором дополнительно оценивался не только уровень знаний по профилю специальности, но и уровень грамотности речи, а так же выяснялся характер мотивации к обучению;

- четвертый этап предполагал прогнозирование и был реализован в виде системы поддержки принятия решений, обеспечивающей:

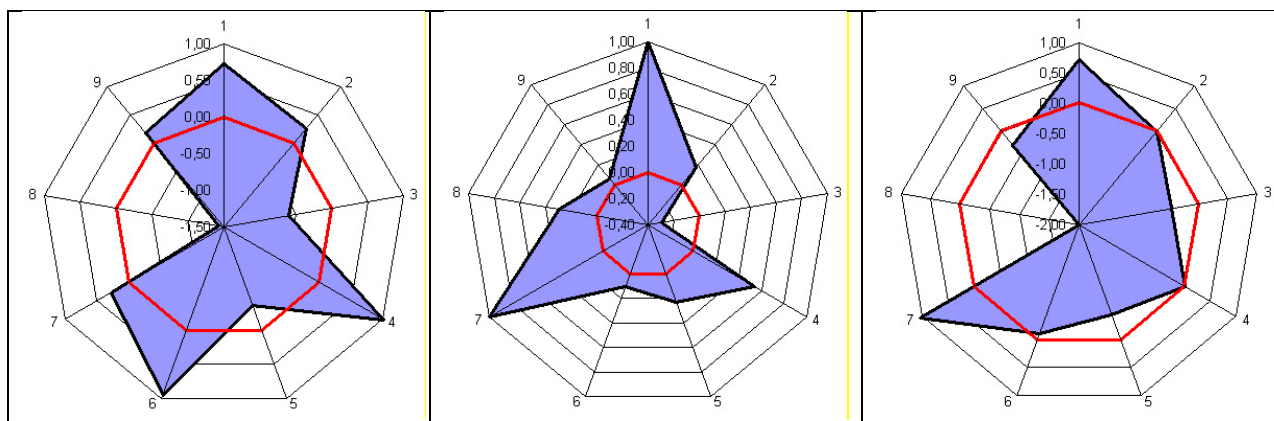


Рис 3. Оценка уровня развития компетентностей трех абитуриентов 2017 года по [3]⁹

1. Оценку облака прецедентов, используя метод экспертных систем на прецедентах [4]: для каждого прецедента рассчитывалось его положение в гиперпространстве факторов, представленного плоскостью осей с оценками векторов априорных и апостериорных параметров модели абитуриента.

2. Визуализация облака оценок прецедентов осуществлялась по указанным осям, при этом успешно прошедшие обучение магистранты отмечались одним цветом (на рис. 4. это оранжевые квадраты), а не успешные – другим (голубые ромбы).

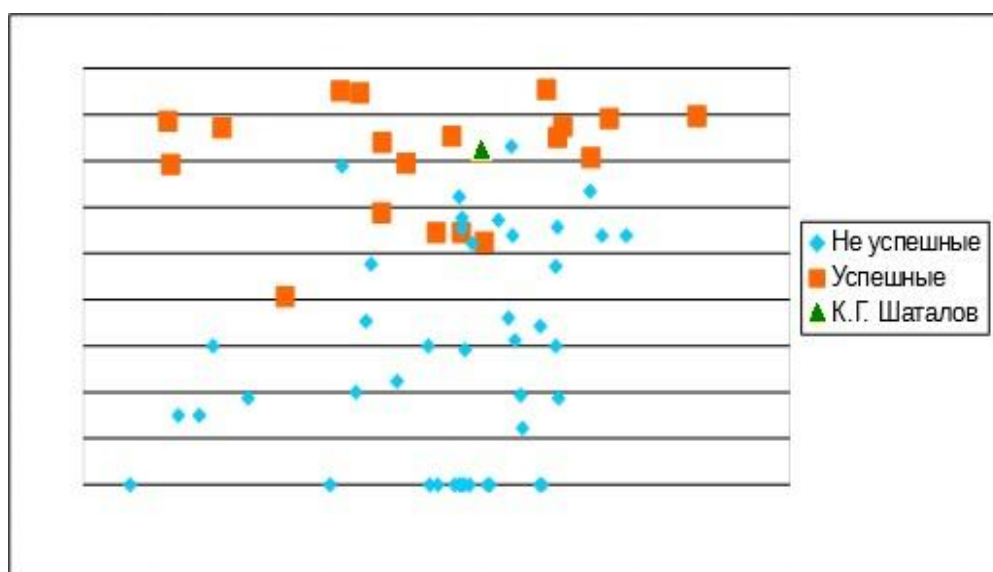


Рис 4. Облако прецедентов для специальности 09.04.01.03 [1]¹⁰

⁹ В основу оценки лег алгоритм экспертного оценивания, базирующийся на критерии достоверности Шортлиффа (Shortliffe E.) и Бьюкенена (Buchanan B.). Оси отображают пронормированные оценки по следующим компетенциям (краткие формулировки): (1) способен самостоятельно работать; (2) способен анализировать и обобщать информацию; (5) способен применять информационные технологии; (7) способен планировать время и работы; (8) способен понимать и составлять алгоритмы; а так же таким характеристикам, как (3) обладать широким кругозором; (4) иметь predisposition к научной работе; (6) иметь интерес к работе в ракетно-космической отрасли; (9) объективно оценивать сложность предстоящей работы.

¹⁰ В основу оценки лег алгоритм экспертного оценивания, базирующийся на критерии достоверности Шортлиффа (Shortliffe E.) и Бьюкенена (Buchanan B.).

3. Прогнозирование потенциального поведения очередного абитуриента на основании схожести (близости) его координат по отношению к ближайшим успешным и не успешным прецедентам. Очевидно, что прогноз будет иметь качественный характер, относящий абитуриента к категории «не будет учиться сразу», «отчислится в ходе обучения», «не выйдет на защиту диссертации», «успешно завершит обучение», дополненный значениями координат в гиперпространстве.

4. Упорядочивание всех абитуриентов, чьи параметрические модели были «наложены» на облако прецедентов, по категории оценки с сортировкой по априорной координате внутри каждой категории.

5. Формулирование рекомендаций и оценка рисков по каждому абитуриенту, опираясь на проблемы, имеющиеся у «похожих» прецедентов из ранее обучавшихся на специальности, а так же опираясь на выявленные закономерности (используется механизм экспертных систем на прецедентах) [5].

Пункт три может быть применен и для магистрантов, уже проходящих обучение на кафедре на любом из этапов подготовки. В этом случае вектор апостериорных параметров будет частично известен (на рис. 4 такая точка показана в виде зеленого треугольного маркера для одного из авторов статьи). В принципе, раскраска маркеров может быть детализирована в разрезе семестров.

Система поддержки принятия решений в различных стадиях готовности применялась для прогнозирования результатов обучения абитуриентов 2015-2017 годов. В качестве примера на рис. 4 приведено облако прецедентов магистрантов для специальности 09.04.01.03, сформированное за 6 лет работы направления. В таблице показаны данные прогноза и его совпадения с реальностью для наборов 2015 и 2016 годов. Видно, что для первого набора совпадение прогноза имело место на 90%, а для второго результаты уже подтвердились на 30% и будут уточняться к моменту защиты магистрантами диссертаций в 2018 году.

Методика, как и средства автоматизации, нуждаются в дальнейшем совершенствовании. В качестве перспективы их развития следовало бы расширить возможности психологической диагностики и уточнить параметрическую модель. Возможно, при организации принятия решений (прогноза) следует применять не только алгоритмы экспертного оценивания по прецедентам, но и методы искусственных нейронных сетей.

Таблица

Прогноз результативности обучения в магистратуре абитуриентов 2015/2016

Абитуриенты 2015 года			Абитуриенты 2016 года		
Инициалы абитуриента	Ближайший прецедент	Фактическое совпадение	Инициалы абитуриента	Ближайший прецедент	Фактическое совпадение ¹¹
А.С.О.	успешный	Нет	Е.С.Н.	успешный	-
А.В.М.	успешный	Да	А.И.А.	близки оба	-

¹¹ Для абитуриентов 2016 года прогноз строился на момент окончания первого курса магистратуры, поэтому совпадение проставлено только для тех магистрантов, которые уже оказались отчислены.

А.Н.Ча.	успешный	Да	Г.А.Г.	близки оба	-
А.А.К.	успешный	Да	М.В.Д.	успешный	-
А.Н.Чу.	не успешный	Да	Р.Б.В.	не успешный	Да
А.Г.К.	не успешный	Да	А.Э.Д.	успешный	-
В.В.К.	не успешный	Да	А.А.Р.	успешный	-
К.Г.Ш.	успешный	Да	А.А.С.	не успешный	-
М.С.К.	не успешный	Да	В.С.Х.	не успешный	Да
И.Ю.Н.	успешный	Да	Т.В.Я.	не успешный	Да
Н.Е.Я.	не успешный	Да	-	-	-

В заключение следует отметить, что предложенный подход не позволяет осуществить зачисление только «перспективных» студентов на указанную специальность, в связи с низким конкурсом. Но это позволяет более внимательно относиться к распределению тем диссертаций и временных ресурсов кафедры при организации индивидуальной работы с магистрантами. Как следствие, прогнозирование результатов обучения на этапе зачисления становится оправданным, а процесс подготовки более управляемым.

Список литературы

11. Шаталов К.Г., Углев В.А. Методическое обеспечение работы приемной комиссии магистратуры на основе анализа статистики обучения // Наука и инновации в XXI веке: Материалы IV международной научно-практической конференции. В 3 ч. Ч. 3. – Пенза, 2017. – С.129–132.

12. Горбань А.Е., Зиновьев А.Ю., Питенко А.А. ViDaExpert: программа для нелинейной визуализации и анализа многомерных данных // Робототехника и искусственный интеллект: Материалы VI Всероссийской научно-технической конференции с международным участием. – Красноярск: Центр информации, 2014. – С. 120-126.

13. Углев В.А., Добронец Б.С. Методика автоматизированного измерения и оценки уровня развития компетентностей // Информатика и образование. – № 2. – 2017. – С. 61-65.

14. Джексон П. Введение в экспертные системы. – М.: Вильямс, 2001. – 624 с.

15. Гаврилова Т.А., Хорошевский В.Ф. Базы знаний интеллектуальных систем. – СПб.: Питер, 2001. – 384 с.

**АНАЛИЗ ИНИЦИАТИВНЫХ НАУЧНЫХ ПРОЕКТОВ ПО ОБЛАСТИ ЗНАНИЯ
"МАТЕМАТИКА, ИНФОРМАТИКА, МЕХАНИКА", ПОДДЕРЖАННЫХ
РОССИЙСКИМ ФОНДОМ ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ**
Р.А. Чиженкова

Институт биофизики клетки РАН, *chizhenkova@mail.ru*

Введение

В настоящее время существует четкое понимание, что наука является реальным базисом политического и экономического процветания государств [2-5, 9-11]. На основе этого развилась усиленная потенция бюджетных расходов на проведение научных исследований [4, 9], что привело к созданию определенных фондов в развитых странах. Что касается России, то одним из первых нововведений молодой РФ явилась организация Российского Фонда Фундаментальных Исследований (РФФИ), которая состоялась в 1992 г. [1, 2]. В открытых публикациях не был сделан анализ научных направлений проектов, поддержанных Фондом, что послужило причиной проведения наших библиометрических исследований [12-15].

Лидирующим аспектом деятельности РФФИ является финансовая поддержка инициативных научных проектов, т.е. реальных исследований ученых [13-15]... Именно эти исследования представляют основу фундаментальной науки, успехи которой в дальнейшем приводят к познанию окружающего нас мира, и дают возможность развития прикладных наук.

За 20 лет своей деятельности РФФИ оказал финансовую поддержку 58667 инициативным, т.е. непосредственно исследовательским, проектам [15]. Поддержанные проекты касались восьми областей знания. По общим количественным аспектам среди них достоверно лидировали проекты по областям знания "Физика, астрономия" (21.10% от общего числа) и "Биология и медицинская наука" (20.61% от общего числа) без существенных различий между собой [15]. На следующем уровне количественных показателей находились научные проекты по трем областям знания: "Математика, информатика, механика"; "Химия и науки о материалах" и "Науки о Земле" [15].

Суммарные материалы по результатам конкурсов инициативных научных проектов по восьми областям знания уже были опубликованы в наших работах [13, 15]. Настоящие исследования посвящены библиометрическому рассмотрению проектов по области знания "Математика, информатика, механика", поддержанных РФФИ в течение 20-летнего периода.

Материалы и методы

Представленные здесь сведения основываются на материалах, опубликованных в Информационных бюллетенях (ИБ) РФФИ, выходящих раз в год и освещающих итоги прошедшего конкурса, которые позволяют рассмотреть количественные данные его результатов по разным областям знания.

Для каждого года составлялась обширная таблица с материалами конкурса проектов по области знания "Математика, информатика, механика" относительно чисел выделенных грантов и поданных заявок. На основе этого вычислялась процентная вероятность поддержки проектов. Проводили объединение результатов по анализируемым рубрикам за весь временной период и вычисляли статистическую значимость различия величин, составляющих эти совокупности. Для статистического анализа использовали сравнение двух выборочных долей вариантов. Кроме того, применяли корреляционный анализ.

Результаты и обсуждение

Число поддержанных РФФИ инициативных проектов по области знания "Математика, информатика, механика" за 20 лет составляло 8209, что представляло 13.99% от общего числа по всем инициативным проектам - 58667. Динамика чисел соответствующих грантов за 20-летний период отражена в табл. 1.

Таблица 1

Динамика чисел поддержанных инициативных проектов по области знания "Математика, информатика, механика" в течение 20-летнего периода

Годы	Инициативные проекты			
	"Математика, информатика, механика"	% от суммы	Все области знания	% от всех областей знания
1993	478	5.82	2845	16.80
1994	456	5.55	2856	15.97
1995	498	6.07	2674	18.62
1996	800	9.75	4742	16.87
1997	392	4.76	2637	14.87
1998	433	5.27	2971	14.57
1999	686	8.36	3132	21.90
2000	379	4.62	2809	13.49
2001	464	5.65	3058	15.17
2002	486	5.92	3046	15.96
2003	336	4.09	2612	12.86
2004	326	3.97	2543	12.82
2005	437	5.32	2936	14.88
2006	296	3.61	3022	9.79
2007	264	3.21	2610	10.11
2008	333	4.06	2947	11.30
2009	279	3.40	2801	9.96
2010	253	3.08	2738	9.24
2011	307	3.74	2922	10.51
2012	306	3.73	2766	11.06
Сумма	8209	100	58667	13.99
Среднее	410.45	5.00	2933.35	13.99

Выделенных грантов приходилось на один год от 253 до 800 и в среднем равнялось 410.45. Из табл. 1 следует, что наибольшие числа поддержанных проектов рассматриваемого направления присутствовали в первое десятилетие при наличии

пиковых значений в 1996 и 1999 гг.. Тогда же отмечались и их весьма выраженные колебания.

Число полученных заявок в ИБ указано только за годы 1996-2000 и 2002-2012. За эти годы было получено 17476 заявок относительно инициативных проектов по области знания "Математика, информатика, механика", что составляет 11.64% от их общего числа по всем областям знания - 150189. Динамика чисел поданных заявок по области знания "Математика, информатика, механика" за исследуемый период представлена в табл. 2.

Таблица 2

Динамика чисел поданных заявок по инициативным проектам в области знания "Математика, информатика, механика" в течение рассматриваемого периода

Годы	Инициативные проекты.			
	Математика, информатика, механика	% от суммы	Все области знания	% от всех областей знания
1993				
1994				
1995				
1996	2222	12.71	15722	14.13
1997	1394	7.98	11129	12.53
1998	1158	6.63	9633	12.02
1999	1291	7.39	9968	12.95
2000	929	5.32	8169	11.37
2001				
2002	1277	7.31	9056	14.10
2003	939	5.37	8000	11.74
2004	911	5.21	7839	11.62
2005	1103	6.31	8475	13.01
2006	831	4.76	9149	9.08
2007	775	4.43	7963	9.73
2008	951	5.44	8530	11.15
2009	872	4.99	8752	9.96
2010	835	4.78	9066	9.21
2011	996	5.70	9658	10.31
2012	992	5.68	9080	10.93
Сумма	17476	100	150189	11.64
Среднее	1092.25	6.25	9386.81	11.64

Числа поданных заявок за один год колебались от 775 до 2222 при среднем значении 1092.25. Табл. 2 показывает, что в начале рассматриваемого периода числа

поданных заявок были преимущественно выше с пиковым значением в 1996 г., а также претерпевали наиболее выраженные колебания.

Описанные здесь количественные характеристики выделенных грантов и представленных заявок позволили определить вероятность поддержки посылаемых материалов. В течение 16 лет, по которым в ИБ даны числа заявок, из 17476 заявок по области знания "Математика, информатика, механика" было поддержано 6313 проекта, что составило 36.12%. Динамика процентных чисел проектов по годам продемонстрирована в табл. 3.

Таблица 3

Динамика процентных чисел поддержанных инициативных проектов по области знания "Математика, информатика, механика" в течение рассматриваемого периода

Годы	Инициативные проекты Математика, информатика, механика		
	Гранты	Заявки	% поддержанных
1993			
1994			
1995			
1996	800	2222	36.00
1997	392	1394	28.12
1998	433	1158	37.39
1999	686	1291	53.14
2000	379	929	40.80
2001			
2002	486	1277	38.06
2003	336	939	35.78
2004	326	911	35.78
2005	437	1103	39.62
2006	296	831	35.62
2007	264	775	34.06
2008	333	951	35.02
2009	279	872	32.00
2010	253	835	30.30
2011	307	996	30.82
2012	306	992	30.85
Сумма	6313	17476	36.12
Среднее	394.56	1092.25	36.12

Годовые значения процентных чисел поддержанных проектов по области знания "Математика, информатика, механика" за данный период составляли от 28.12 до 53.14 при средней величине 36.12. Как видно из рис. 3, пикового значения процентное число

поддержанных проектов достигало в 1999 г.. Кроме того, наибольшие колебания процентных чисел поддерживаемых проектов приходились на начальную часть 16-летнего периода.

Таким образом, за 20 лет своей деятельности РФФИ оказал значительную поддержку проведению исследований в области знания "Математика, информатика, механика". В XXI-ом веке ожидается радикальное изменение парадигмы в математике, которое отразится на всех областях знания [8]. При этом займет должное место информационно-теоретический поворот в интерпретации квантовой механики, начавшийся в конце XX-ого века, что в той или иной мере затронет все области знания [7]. И то, и другое создаст новую основу для познания вселенной [6]. В настоящей работе проанализированы числа выделенных грантов, числа подаваемых заявок и процентные числа поддерживаемых проектов по области знания "Математика, информатика, механика", что позволило выявить следующие моменты.

Во-первых, число поддерживаемых инициативных проектов по области знания "Математика, информатика, механика" достоверно было значительно меньше, чем по областям знания "Биология и медицинская наука" и "Физика, астрономия": соответственно в 1.47 раз ($p < 0.01$ при $U = 29.97$) и 1.51 раз [15] ($p < 0.01$ при $U = 32.20$) [15]. Тем не менее, оно представляло весьма существенную величину - 8209, что составляло 13.99% от общего числа по всем восьми рассматриваемым областям знания. Доля поданных заявок проектов по области знания "Математика, информатика, механика" в общем числе таковых обладала близким показателем - 11.64%. Процентное число поддерживаемых проектов по области знания "Математика, информатика, механика" равнялось 36.12, что достоверно превышало аналогичное процентное число в суммарных данных по всем областям знания - 31.45 [15] ($p < 0.01$ при $U = 12.26$).

Во-вторых, в течение рассматриваемого периода отмечались существенные колебания годовых значений чисел выделенных грантов, чисел подаваемых заявок и процентных чисел поддерживаемых проектов. Наблюдалось некоторое сходство динамик чисел грантов и чисел заявок и выраженное их отличие от динамики процентных чисел поддерживаемых проектов. Тем не менее, во всех трех случаях колебания соответствующих величин отмечались преимущественно в начальной части анализируемого периода. При этом колебания чисел грантов были в 3.16 раз, сделанных заявок - 2.87 раз и процентов поддерживаемых проектов - в 1.89 раз.

На основе корреляционного анализа было установлено наличие некоторого параллелизма чисел выделенных грантов и чисел подаваемых заявок и ($r = 0.89$; $p < 0.01$), что свидетельствует о значительной роли общего состояния внутренней возможности в поддержке научных проектов РФФИ. Кроме того, была выявлена положительная корреляционная взаимосвязь чисел выделенных грантов и чисел подаваемых заявок по области знания "Математика, информатика, механика" с соответствующими величинами суммарных данных по всем областям знания в конкурсе инициативных проектов ($r = 0.75$ и 0.93 ; $p < 0.01$) [15]. Однако среди процентных чисел поддерживаемых проектов такой эффект отсутствовал. ($r = 0.43$; $p > 0.05$).

Заключение

Очень своевременным мероприятием только-что возникшей РФ явилось создание РФФИ для поддержки работы научных коллективов и отдельных ученых. За два десятилетия деятельности РФФИ была оказана существенная поддержка проведению фундаментальной исследований, в том числе и по области знания "Математика, информатика, механика". Для повышения возможности поддержки научных проектов весьма желательно увеличение финансового наполнения Фонда. Следует отметить, что российское правительство в ближайшее время намерено увеличить ресурсное наполнение Фонда. Об этом было сказано во время выступления В.В. Путиным на

Общем собрании Российской академии наук в 2012 г., где он дал высокую оценку работе Фонда [10].

Список литературы

1. Алфимов М.В., Минин В.А., Либкинд А.Н. Страна наука - РФФИ // Вестник РФФИ. - 2000. - № 2(20). - С. 5-29.
2. Арутюнов В.С. Наука как один из важнейших институтов современного государства // В: Наука России. От настоящего к будущему / Ред В.С. Арутюнов, Г.В. Лисичкин, Г.Г. Малинецкий. - М.: Книжный дом "ЛИБРОКОМ", 2009. - С. 9-29.
3. Ефремов Ю.Н. Зачем нужна наука миру и России // Вестник РФФИ. - 2000. - № 1(19). - С. 40-43.
4. Коннов В.И. Самоуправление на "передовой": становление национального научного фонда США // Вестник РФФИ. - 2007. - №4(54). - С. 10-15.
5. Лебедев С.А. Праксиология науки // Вопросы философии. - 2012. - №4. - С. 52-63.
6. Ддойд С. Программируя вселенную. Квантовый компьютер и будущее науки / Пер. с англ. - М.: Альпина нон-фикшн, 2013. - 256 с.
7. Мамчур Е.А. Информационно-теоретический поворот в интерпретации квантовой механики: философско-методологический анализ // Вопросы философии. - 2014. - №1. - С. 57-82.
8. Нариньяни А.С. Математика XXI - радикальная смена парадигмы. Модель, а не алгоритм // Вопросы философии. - 2011. - №1. - С. 71-82.
9. Обама Б. Выступление в Национальной академии наук 27-ого апреля 2009 г. // В защиту науки. Бюл. № 6. Ред. Э.П. Кругляков. - М.: Наука, 2009. - С. 185-198.
10. Путин В.В. Выступление на Общем собрании Российской академии наук 22 мая 2012 // В защиту науки. Бюл. № 11. Ред. Э.П. Кругляков. - М.: Наука, 2012. - С. 7-14.
11. Чиженкова Р.А. Динамика нейрофизиологических исследований действия неионизирующей радиации во второй половине XX-ого века. - М.: Издат. дом Акад. Естествознания, 2012. - 88 с.
12. Чиженкова Р.А. Библиометрический анализ научных проектов, поддержанных Российским Фондом Фундаментальных Исследований за 20 лет: виды конкурсов // Успехи современного естествознания. - 2014. - №5(часть 2) - С. 145-150.
13. Чиженкова Р.А. Библиометрический анализ инициативных научных проектов по разным областям знания, поддержанных Российским Фондом Фундаментальных Исследований за 20 лет // Международный журнал прикладных и фундаментальных исследований. - 2014. - №7. - С. 100-105.
14. Чиженкова Р.А. Российский фонд фундаментальных исследований. Библиометрический анализ поддержки научных проектов. - Germany: Palmarium Academic Publishing, 2015. - 88 с.
15. Чиженкова Р.А. Российский фонд фундаментальных исследований: Библиометрический анализ инициативных научных проектов. - М.: Изд. дом Акад. Естествознания, 2016. - 82 с.

ПРИНЦИПЫ ПОСТРОЕНИЯ НЕЙРО-НЕЧЁТКОЙ СИСТЕМЫ ТЕРАПИИ ЛЕКАРСТВЕННЫМИ СРЕДСТВАМИ¹²

О.И. Ширяева

Институт информационных и вычислительных технологий Комитета науки МОН РК
oshiryayeva@gmail.com

Цель доклада заключается в формулировке принципов построения нейро-нечёткой системы терапии организма лекарственными средствами на примере разработки программного обеспечения обучения и моделирования гибридной нейро-нечёткой сети искусственной иммунной системы терапии организма сульфаниламидами.

Для разработки базы данных входных и выходных переменных терапии заболеваний различными медицинскими препаратами, составим экспертную таблицу (табл. 1). К основным переменным, влияющим на проведение терапии, отнесём следующие:

- свойства лекарственного препарата;
- лабораторные показатели болезни;
- свойства организма;
- терапевтические дозы лекарственных препаратов.

Каждой из переменной соответствует лингвистическое наименование, обозначение, единица измерения и нечёткие множества [1], представленные степенью принадлежности, μ . Все переменные, и соответствующие им нечёткие множества, формируются на основе экспертного мнения специалистов.

Таблица 1

Представление входных и выходных переменных

	Лингвистическая переменная	Нечёткие множества		
ВХОДНЫЕ ПЕРЕМЕННЫЕ				
Свойства лекарственного препарата				
$\mu(x)$	– «лингвистическое наименование», – обозначение, – единица измерения	$\mu_{x1}(x)$	$\mu_{x2}(x)$	...
Лабораторные показатели болезни				
$\mu(y)$	– «лингвистическое наименование», – обозначение, – единица измерения	$\mu_{y1}(y)$	$\mu_{y2}(y)$	...
Свойства организма				
$\mu(z)$	– «лингвистическое наименование», – обозначение, – единица измерения	$\mu_{z1}(z)$	$\mu_{z2}(z)$	...
ВЫХОДНЫЕ ПЕРЕМЕННЫЕ				
Терапевтические дозы лекарственных препаратов, $f(u)$ – при системе нечеткого вывода типа Сугено				
$f(u)$	– обозначение, – единица измерения	u_1	u_2	...

В соответствии с полученными результатами исследований [2], система

¹² Исследования проводятся по гранту №ГР 0115РК00549 МОН РК по теме: Компьютерный молекулярный дизайн лекарственных препаратов на основе иммуносетевого моделирования (2015-2017 гг.).

нечёткого вывода имеет восемь входных переменных, одну выходную (табл. 2).

Таблица 2

Представление входных переменных системы терапии пиелонефрита

	Лингвистическая переменная	Нечёткие множества термов			
ВХОДНЫЕ ПЕРЕМЕННЫЕ					
СВОЙСТВА СУЛЬФАНИЛАМИДОВ					
$\mu(x_1)$	«длительность выведения из организма сульфаниламидов», T(1/2), час	$\mu_{11}(x_1)$	$\mu_{12}(x_1)$	$\mu_{13}(x_1)$	$\mu_{14}(x_1)$
		кратковрем. действия	средней длит.	длит. действия	сверхдлит
		8 часов	8÷20	24÷48	65÷120
$\mu(x_2)$	«ацетилирование сульфаниламидов», А, %	$\mu_{21}(x_2)$	$\mu_{22}(x_2)$		
		медленные ацетиляторы	быстрые ацетиляторы		
		0÷50	50÷100		
ЛАБОРАТОРНЫЕ ПОКАЗАТЕЛИ ПИЕЛОНЕФРИТА					
А – фаза активного воспалительного процесса; Л – фаза латентного воспалительного процесса; Р – фаза ремиссии, или клинического выздоровления.					
$\mu(x_3)$	«фазы хронического пиелонефрита в зависимости от количества лейкоцитов», QI*1000,ед. в 1 мл	$\mu_{31}(x_3)$	$\mu_{32}(x_3)$	$\mu_{33}(x_3)$	
		Р	Л	А	
		0	<25 000	>25 000	
$\mu(x_4)$	«в зависимости от количества микроорганизмов», Qm *1000, ед в 1 мл	$\mu_{41}(x_4)$	$\mu_{42}(x_4)$	$\mu_{43}(x_4)$	
		Р	Л	А	
		0	0÷10 000	>100 000	
$\mu(x_5)$	«в зависимости от количества активных лейкоцитов», QAL , %	$\mu_{51}(x_5)$	$\mu_{52}(x_5)$	$\mu_{53}(x_5)$	
		Р	Л	А	
		0	15÷30	>30	
$\mu(x_6)$	«в зависимости от СОЭ», V , мм/ч	$\mu_{61}(x_6)$	$\mu_{62}(x_6)$	$\mu_{63}(x_6)$	
		Р	Л	А	
		<12	≈12	>12	
СВОЙСТВА ОРГАНИЗМА					
$\mu(x_7)$	«возраст», Количество лет	$\mu_{71}(x_7)$	$\mu_{72}(x_7)$	$\mu_{73}(x_7)$	$\mu_{74}(x_7)$
		"дети"	"мол.люди"	"взрослый"	"пожилой"
		0÷14	14÷25	21÷65	>60
$\mu(x_8)$	«Масса» Центили, %	$\mu_{81}(x_8)$	$\mu_{82}(x_8)$	$\mu_{83}(x_8)$	
		"низкий"	"норма"	"высокий"	
		0÷25	10÷90	75÷100	

Для терапии сульфаниламидами такого заболевания, как пиелонефрита, в отчете по проекту по данной теме, в 2016 году [2], рассмотрены различные сульфаниламидные

препараты: уросульфам, этазол, сульфапиридазин, сульфамонометоксин, сульфадиметоксин, норсульфазол (сульфатиазол), сульфадимезин (сульфадимидин), этазол (сульфаэтидол), сульфапиридазин (сульфаметоксипиридазин, кинекс, спофадазин), сульфамонометоксин (дуфадин), сульфадиметоксин (мадрибон), сульфален (келфизин), бактрим (бисептол, септрим, орибакт, ко-тримоксазол).

Каждый из препаратов отличается спецификой использования при различных показаниях. Для каждого препарата, при построении системы нечёткого вывода, необходимо составлять базу правил стратегий терапии пиелонефрита с использованием сульфаниламидных препаратов. В качестве примера, выбран вариант терапии организма уросульфамом (табл. 3).

Таблица 3

Представление выходных переменных системы терапии пиелонефрита уросульфамом

ВЫХОДНАЯ ПЕРЕМЕННАЯ, $f(u)$ – дозы уросульфана									
ПДЛ – пороговая доза лекарства; СДЛ – средняя доза лекарства; ВДЛ – высшая доза лекарства; ПСЛ – пороговая смертельная доза.									
$f(u)$ «терапевтические дозы уросульфана», р, гр	u_1		u_2		u_3		u_4		
	1	2	2	4	4	7	7	8	
	ПДЛ		СДЛ		ВДЛ		ПСЛ		

Базе правил стратегий терапии пиелонефрита соответствует двенадцать правил [1], в зависимости от входных переменных (табл. 2).

Основными недостатками экспертных систем, основанных на нечеткой логике, являются, прежде всего, субъективность выбора набора правил и параметров функций принадлежности входных переменных, которая может серьезно сказаться на результате, и сложность корректировки этих параметров на основе экспериментальных данных. С этими проблемами позволяют бороться нейро-нечеткие системы, объединяющие в себе лучшие качества нейронных сетей и нечетких систем (обучение нейронной сети называется процесс идентификации параметров сети на основе набора входо-выходных данных - обучающего множества). Свойство обучаемости нейронных сетей является одной из главных причин их повсеместного использования.

Есть несколько способов объединения нечеткой логики и нейронных сетей. Наибольшую область применения нашли гибридные нейро-нечеткие системы. Их характерной чертой является то, что они всегда могут быть рассмотрены как системы нечетких правил, при этом настройка функций принадлежности в предпосылках и заключениях правил на основе обучающего множества производится с помощью нейронной сети.

Существует несколько архитектур гибридных систем, но далее будет использована архитектура адаптивной нейро-нечеткой системы вывода ANFIS (Adaptive Neuro-Fuzzy Inference System,) которая реализует систему нечеткого вывода Сугено в виде нейронной сети прямого распространения сигнала.

Загрузка данных нейро-нечеткой сети терапии сульфаниламидами производится с внешнего сформированного файла `gibrid_medicine.txt` (Рис. 1). Нечеткая модель гибридной системы содержит 8 входных переменных и одну выходную – соответствующие обучающие данные сведены в отдельную таблицу.

gibrid_medicine - Notepad									
File	Edit	Format	View	Help					
65	50	50000	75000	50	12.5	50	50	2.35	
65	60	60000	85000	60	13	60	50	2.19	
65	28.18	1.545e+04	1.432e+05	40.91	22.05	28.18	79.09		
47.32;9.2;9.167e+04;1.98e+04;10.8;4.3;83.73;79.09 3									

Рис. 1. Обучающие данные для построения модели гибридной сети

После загрузки файла с обучающими данными в редактор ANFIS, экспериментальные данные и результаты моделирования выводятся в виде множества точек в двумерном пространстве (Рис. 2). При этом по оси абсцисс откладывается порядковый номер строки данных в обучающей выборке, а по оси ординат - значение выходной переменной, u , для данной строки выборки. Используются следующие маркеры:

- голубая окружность (o) – обучающая выборка;
- красная звездочка (*) – результаты моделирования.

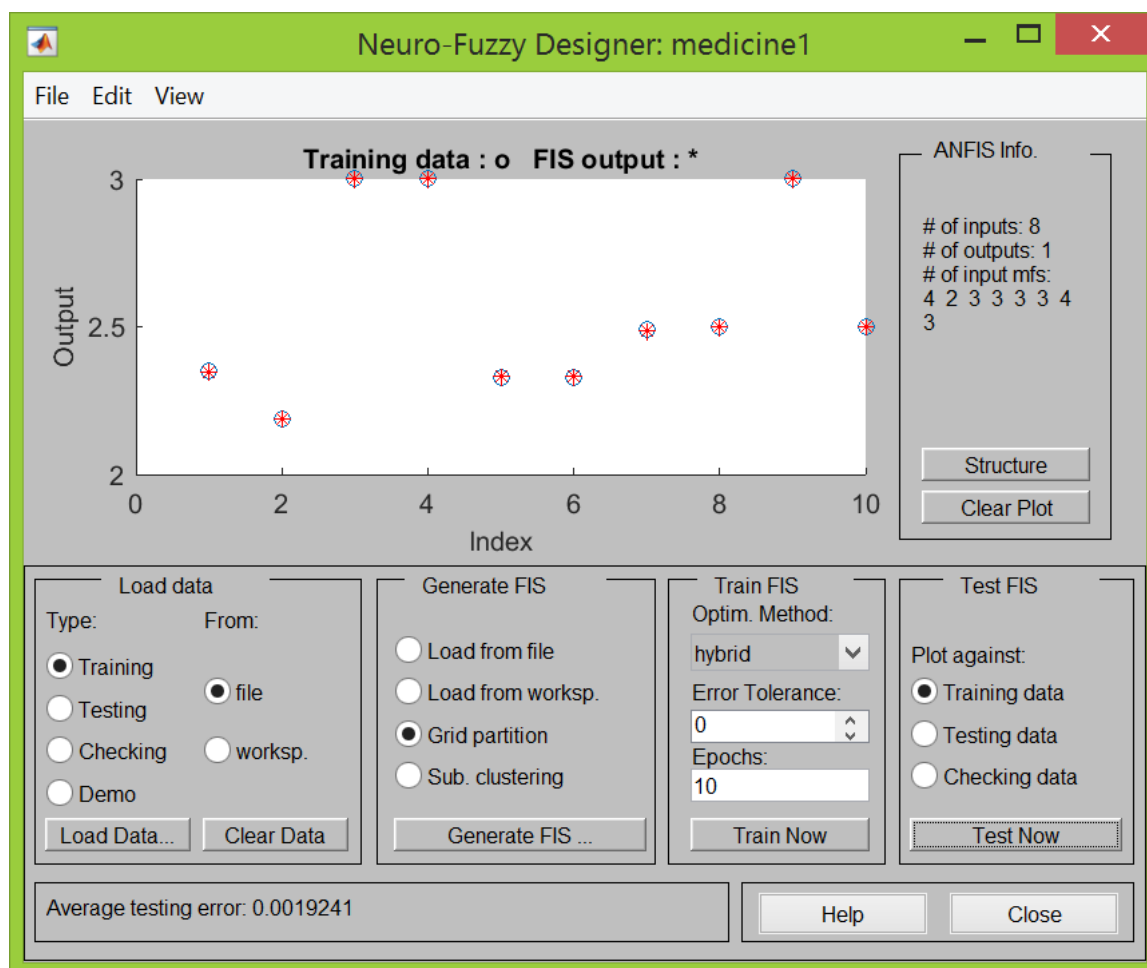


Рис. 2. Графический интерфейс редактора ANFIS после загрузки обучающих данных

Перед генерацией структуры системы нечёткого вывода Сугено, после вызова диалогового окна свойств, задаётся для каждой из входных переменных

соответствующее число лингвистических термов (табл. 2), а в качестве типа их функций принадлежности выбираются функции, разработанные в проекте [2]. Выходные значения принимают вид функций в соответствии с типом нечёткого вывода Сугено. Выходные значения, в соответствии с табл. 3, имеют линейный вид функции активации нейрона.

После обучения гибридной сети можно визуально оценить структуру построенной нечёткой модели (Рис.3).

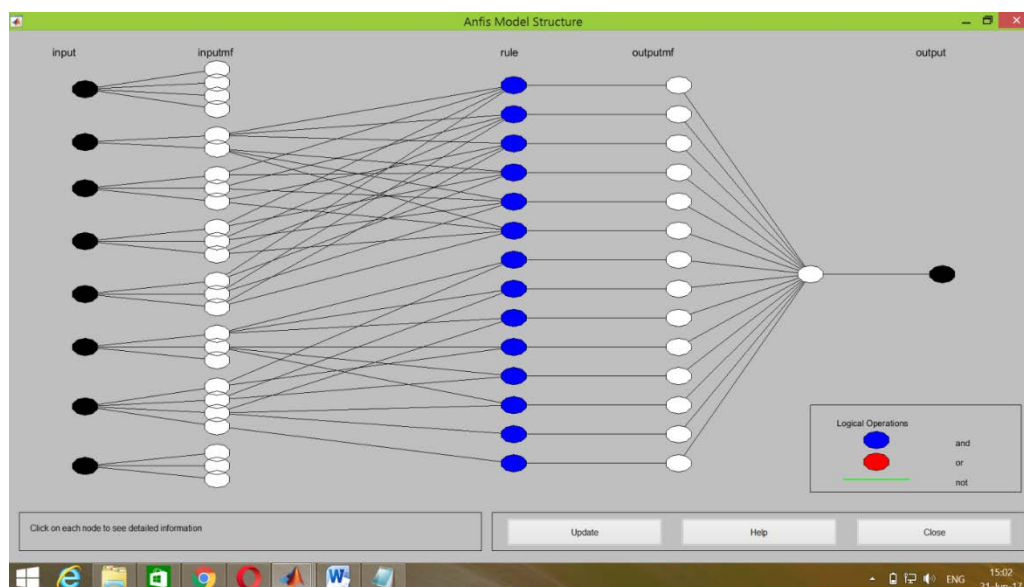


Рис.3. Структура сгенерированной гибридной сети в редакторе AnfisModelStructure

Для исследования построенной модели гибридной сети можно воспользоваться программой просмотра правил – Rule Viewer (Рис.4).



Рис.4. Интерфейс просмотра правил в RuleEditor сгенерированной системы

Для получения значения выходной переменной необходимо задать соответствующие значения входных переменных, в соответствии с рекомендациями

нечёткого вывода. На графике функций принадлежности выходной переменной указывается искомое значение выходной переменной ($u = 2.35$).

Заключение. Приведены результаты разработки принципов построения нейро-нечёткой системы терапии организма лекарственными средствами. Полученные результаты апробированы в ходе разработки программного обеспечения обучения и моделирования гибридной нейро-нечёткой сети искусственной иммунной системы терапии организма сульфаниламидами. Программное обеспечение реализовано в среде MATLAB с использованием редактора адаптивной системы нейро-нечеткого вывода ANFIS.

Список литературы

1. Shirayeva O.I. Investigation of artificially immune system with using of fuzzy logic / Shirayeva O.I., Denisova T.G. // Новосибирск: Вычислительные технологии (совместный сборник журналов с «Вестник КазНУ. Серия математика, механика, информатика»). — 2015. — С. 209 – 217.
1. Keles A., Keles A., Yavuz U. Expert System Based on Neuro-Fuzzy Rules for Diagnosis Breast Cancer / Keles A., Keles A., Yavuz U. // Expert System with Applications. — 2011. — N38. — pp. 5719-5726.

СПИСОК ДОКЛАДОВ

М.М. Алкезуини, В.И. Горбаченко УСКОРЕННЫЕ МЕТОДЫ ПЕРВОГО ПОРЯДКА ОБУЧЕНИЯ СЕТЕЙ РАДИАЛЬНЫХ БАЗИСНЫХ ФУНКЦИЙ	3
С.С. Анцыферов, К.Н. Фазилова, К.Е. Русанов ОПРЕДЕЛЕНИЕ ПОКАЗАТЕЛЕЙ УСТОЙЧИВОГО ФУНКЦИОНИРОВАНИЯ НЕЙРОПОДОБНЫХ СИСТЕМ	8
Д.Д.Барышев, О.И.Пятковский ФОРМИРОВАНИЕ ПЛАНА ЗАКУПОК АВТОМОБИЛЬНЫХ ЗАПЧАСТЕЙ НА ОСНОВЕ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ И ГЕНЕТИЧЕСКИХ АЛГОРИТМОВ	14
В.В.Бирюков, С.В.Новикова, Ю.А. Путинцева, М.Г.Садовский, К.В. Крутовский СИММЕТРИЯ В СТАТИСТИЧЕСКИХ СТРУКТУРАХ ТРАНСКРИПТОМА И ГЕНОМА ХВОЙНЫХ ВИДОВ	19
Н.М. Богатов, А.Н. Богатова, В.С. Лахин, С.А. Сухих ВЛИЯНИЕ ТЕХНИК САМОРЕГУЛЯЦИИ НА ПОКАЗАТЕЛИ ЗДОРОВЬЯ СТУДЕНТОВ	27
Л.А.Компаниец, О.С.Володько, Л.В.Гаврилова МЕТОД ГЛАВНЫХ И НЕЗАВИСИМЫХ КОМПОНЕНТ В АНАЛИЗЕ НАТУРНЫХ ИЗМЕРЕНИЙ ТЕМПЕРАТУРЫ НЕГЛУБОКОГО ВОДОЕМА	31
Е.В. Попова, М.В. Гунер, О.И. Пятковский ПОСТРОЕНИЕ МОДЕЛИ ОЦЕНКИ И ПРОГНОЗА ПОСЕЩАЕМОСТИ ПРЕДПРИЯТИЙ ОБЩЕСТВЕННОГО ПИТАНИЯ НА ОСНОВЕ ПРИМЕНЕНИЯ НЕЙРОСЕТЕВЫХ ТЕХНОЛОГИЙ	37
Е.Н. Потылицына, Е.В. Сугак ПРОГНОЗИРОВАНИЕ ОЖИДАЕМОЙ ПРОДОЛЖИТЕЛЬНОСТИ ЖИЗНИ НАСЕЛЕНИЯ С ИСПОЛЬЗОВАНИЕМ ИСКУССТВЕННЫХ НЕЙРОННЫХ СЕТЕЙ	43
М.Г.Садовский, А.Н.Остыловский, Е.Ю.Бушмелев О НЕКОТОРЫХ МЕТОДАХ КЛАСТЕРИЗАЦИИ МНОГОМЕРНЫХ ДАННЫХ, УЧИТЫВАЮЩИХ ТОПОЛОГИЮ ПРИБЛИЖАЮЩЕГО ИХ МНОГООБРАЗИЯ МАЛОЙ РАЗМЕРНОСТИ	49
Г.А. Самигулина, А.Т. Нюсупов, А.С. Шаяхметова	

РАЗРАБОТКА МУЛЬТИАГЕНТНОЙ SMART-СИСТЕМЫ ДИСТАНЦИОННОГО ОБУЧЕНИЯ ЛЮДЕЙ С ОГРАНИЧЕННЫМИ ВОЗМОЖНОСТЯМИ ЗРЕНИЯ	54
З.И. Самигулина, Т.И. Самигулин ПРИМЕНЕНИЕ СОВРЕМЕННЫХ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ СИНТЕЗА НОВЫХ ЛЕКАРСТВЕННЫХ ПРЕПАРАТОВ СУЛЬФАНИЛАМИДОВ	59
М.Ю. Сенашова, М.Г. Садовский СТРУКТУРА ГЕНОМОВ ЦИАНОБАКТЕРИЙ	65
М.Ю. Сенашова, М.Г. Садовский АНАЛИЗ КЛАСТЕРНОЙ СТРУКТУРЫ ГЕНОМОВ ХЛОРОПЛАСТОВ ВЫЯВЛЯЕТ ИХ ОТЛИЧИЯ ОТ ГЕНОМОВ ПРОКАРИОТ	70
В.П.Сизиков, В.И. Разумов ТЕОРИИ ВСЕГО И СМЫСЛОВОЕ РАЗВИТИЕ	79
Н.В. Смыкова, О.И. Пятковский, Н.А. Звеков ИСПОЛЬЗОВАНИЕ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОЦЕНКИ ПОДГОТОВКИ СТУДЕНТОВ К ВЫПОЛНЕНИЮ ТРУДОВЫХ ФУНКЦИЙ ПРОФЕССИОНАЛЬНЫХ СТАНДАРТОВ	85
Субботин С.А. СИНТЕЗ МОДЕЛЕЙ ПРИНЯТИЯ РЕШЕНИЙ НА ОСНОВЕ МЕТОДОВ ВЫЧИСЛИТЕЛЬНОГО ИНТЕЛЛЕКТА И ФРАКТАЛЬНОГО АНАЛИЗА	90
В.А.Углев ПРОВЕРКА ГИПОТЕЗЫ ОБ УРОВНЕ РАЗВИТИЯ КОМПЕТЕНТНОСТИ С ИСПОЛЬЗОВАНИЕМ КРИТЕРИЯ ШОРТЛИФФА И БЬЮКЕНЕНА	96
В.А.Углев, К.Г. Шаталов, С.В. Холодилов, В.А. Холодилова СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ПО ОТБОРУ АБИТУРИЕНТОВ В МАГИСТРАТУРУ С ЭЛЕМЕНТАМИ ПРОГНОЗИРОВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ	101
Р.А. Чиженкова АНАЛИЗ ИНИЦИАТИВНЫХ НАУЧНЫХ ПРОЕКТОВ ПО ОБЛАСТИ ЗНАНИЯ "МАТЕМАТИКА, ИНФОРМАТИКА, МЕХАНИКА", ПОДДЕРЖАННЫХ РОССИЙСКИМ ФОНДОМ ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ	107

О.И. Ширяева

ПРИНЦИПЫ ПОСТРОЕНИЯ НЕЙРО-НЕЧЁТКОЙ СИСТЕМЫ
ТЕРАПИИ ЛЕКАРСТВЕННЫМИ СРЕДСТВАМИ

113

СПИСОК АВТОРОВ

Ф.И.О.	Место работы	Стр.
Алкезуини Мухи Мартада	440026, г. Пенза, ул. Красная, д. 40. Пензенский государственный университет E-mail: mohieit@mail.ru	3
Анцыферов Сергей Сергеевич	119454, Россия, г. Москва, пр. Вернадского, 78, Московский технологический университет (МИРЭА) E-mail: c_standard@fel.mirea.ru	8
Барышев Денис Дмитриевич	656038 г. Барнаул, Алтайский государственный технический университет им. И. И. Ползунова, E-mail: denis.baryshev@bk.ru	14
Бирюков Владислав Владимирович	660041, г. Красноярск, пр. Свободный, 79, Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологии E-mail: vladbir2010@gmail.com	19
Богатов Николай Маркович	350040, г. Краснодар, ул. Ставропольская, 149, ФГБОУ ВПО «Кубанский государственный университет», E-mail: bogatov@phys.kubsu.ru	27
Богатова Анна Николаевна	350040, г. Краснодар, ул. Ставропольская, 149, ФГБОУ ВПО «Кубанский государственный университет», т. 8(861)2199501266, E-mail: bogatov@phys.kubsu.ru	27
Бушмелев Евгений Юрьевич	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: msad@icm.krasn.ru	49
Володько Ольга Станиславовна	236022, Калининград, просп. Мира, 1-3, Институт океанологии АО РАН E-mail: olga.pitalskaya@gmail.com	31
Гаврилова Людмила Владимировна	660041, г. Красноярск, пр. Свободный, 79, Сибирский федеральный университет E-mail: lv.gavrilova@gmail.com	31
Горбаченко Владимир Иванович	440026, Пенза, ул. Красная, 40, Пензенский государственный университет E-mail: gorvi@mail.ru	3
Гунер Михаил Викторович	656038, Российская Федерация, Алтайский край, г. Барнаул, пр. Ленина, 46, Алтайский государственный технический университет им. И. И. Ползунова, E-mail: horyzont1@mail.ru	37
Звеков Никита Александрович	656038, Российская Федерация, Алтайский край, г. Барнаул, пр. Ленина, 46, Алтайский государственный технический университет им. И. И. Ползунова E-mail: nzvekov@inbox.ru	85
Компаниец Лидия Алексеевна	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: kla@icm.krasn.ru	31
Крутовский Кон-	119991, ГСП-1, Москва, ул. Губкина, дом 3, Институт об-	19

стантин Валерьевич	щей генетики им. Н.И. Вавилова Российской Академии Наук E-mail: kkrutovsky@gmail.com	
Лахин Владислав Сергеевич	350040, г. Краснодар, ул. Ставропольская, 149, ФГБОУ ВПО «Кубанский государственный университет», E-mail: bogatov@phys.kubsu.ru	27
Новикова Серафима Влпдимировна	660041, г. Красноярск, пр. Свободный, 79, Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологии E-mail: msad@icm.krasn.ru	19
Нюсупов Адлет Талапович	050010, Республика Казахстан, г. Алматы, ул. Пушкина, 125, Институт информационных и вычислительных технологий E-mail: moniumverse@otlook.com	54
Остыловский Анатолий Николаевич	660041, г. Красноярск, пр. Свободный, 79, Институт математики и фундаментальной информатики СФУ (ИМиФИ СФУ) E-mail: hinayana@g-service.ru	49
Попова Екатерина Вячеславовна	656038, Российская Федерация, Алтайский край, г. Барнаул, пр. Ленина, 46, Алтайский государственный технический университет им. И. И. Ползунова, E-mail: horyzont1@mail.ru	37
Потылицына Елена Николаевна	660000, Красноярск, просп. имени газеты Красноярский Рабочий, 31, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева E-mail: Leonova_en@mail.ru	43
Путинцева Юлия Андреевна	660041, г. Красноярск, пр. Свободный, 79, Сибирский федеральный университет, Институт фундаментальной биологии и биотехнологии E-mail: yuliya-putintseva@rambler.ru	19
Пятковский Олег Иванович	656038, Российская Федерация, Алтайский край, г. Барнаул, пр. Ленина, 46, Алтайский государственный технический университет им. И. И. Ползунова E-mail: poi1952@mail.ru	14, 37, 85
Разумов Владимир Ильич	Омский государственный университет им. Ф.М. Достоевского, Россия, 644077, Омск, проспект Мира 55-а E-mail: razumov@omsu.ru	79
Русанов Константин Евгеньевич	119454, Россия, г. Москва, пр. Вернадского, 78, Московский технологический университет (МИРЭА) E-mail: c_standard@fel.mirea.ru	8
Садовский Михаил Георгиевич	660036, Красноярск, Академгородок, Институт вычислительного моделирования СО РАН E-mail: msad@icm.krasn.ru	19, 49, 65, 70
Самигулин Тимур Илдусович	050010, Республика Казахстан, г. Алматы, ул. Пушкина, 125, Институт информационных и вычислительных технологий E-mail: timur.samigulin@yandex.kz	59
Самигулина Галина Ахметовна	050010, Республика Казахстан, г. Алматы, ул. Пушкина, 125, Институт информационных и вычислительных технологий E-mail: galinasamigulina@mail.ru	54

Самигулина Зарина Ильдусовна	050013, г. Алматы, ул.Сатпаева, 22 а, Казахский национальный технический исследовательский университет, Институт информационных и телекоммуникационных технологий E-mail: zarinasamigulina@mail.ru	59
Сенашова Мария Юрьевна	Институт вычислительного моделирования СО РАН, 660036, Красноярск, Академгородок E-mail: msen@icm.krasn.ru	65, 70
Сизиков Виктор Петрович	Омский государственный университет путей сообщения, 644046, Россия, г. Омск, пр. Маркса, 35. Т./ф.: (3812) 31-42-19. E-mail: v_p_sizikov@mail.ru	79
Смыкова Наталья Владимировна	656038, Российская Федерация, Алтайский край, г. Барнаул, пр. Ленина, 46, Алтайский государственный технический университет им. И. И. Ползунова E-mail: n-smykova@mail.ru	85
Субботин Сергей Александрович	Запорожский национальный технический университет, Украина, 69063, Запорожье, ул. Жуковского, 64, т.: +38-(097) 355-61-55, E-mail: subbotin@zntu.edu.ua	90
Сугак Евгений Викторович	660000, Красноярск, просп. имени газеты Красноярский Рабочий, 31, Сибирский государственный университет науки и технологий имени академика М.Ф. Решетнева E-mail: Leonova_en@mail.ru	43
Сухих Станислав Алексеевич	350040, г. Краснодар, ул. Ставропольская, 149, ФГБОУ ВПО «Кубанский государственный университет», E-mail: suhih_52@mail.ru	27
Углев Виктор Александрович	662971, Железногорск, ул. Кирова, 12-а, Железногорский филиал Сибирского федерального университета E-mail: uglev-v@yandex.ru	96, 101
Фазилова Ксения Наильевна	119454, Россия, г. Москва, пр. Вернадского, 78, Московский технологический университет (МИРЭА) E-mail: fazilova@mirea.ru	8
Холодилова Валерия Александровна	662971, Железногорск, ул. Кирова, 12-а, Железногорский филиал Сибирского федерального университета E-mail: daloka@list.ru	101
Холодилов Сергей Валерьевич	662971, Железногорск, ул. Кирова, 12-а, Железногорский филиал Сибирского федерального университета E-mail: aviasergey@yandex.ru	101
Чиженкова Рогнеда Александровна	142290, г. Пущино Московской области, Институт биофизики клетки РАН E-mail: chizhenkova@mail.ru	107
Шаталов Константин Геннадьевич	662971, Железногорск, ул. Кирова, 12-а, Железногорский филиал Сибирского федерального университета E-mail: shatalovkg@gmail.com	101
Шаяхметова Асем Серикбаевна	050010, Республика Казахстан, г. Алматы, ул. Пушкина, 125, Институт информационных и вычислительных технологий, E-mail: asemshayakhmetova@mail.ru	54

Ширяева Ольга Ивановна	050010, Республика Казахстан, Алматы, ул. Пушкина, 125, Институт информационных и вычислительных технологий КН МОН РК E-mail: oshiryayeva@gmail.com	113
---------------------------	--	-----

Научное издание

Нейроинформатика, её приложения и анализ данных

Материалы XXV Всероссийского семинара 29 сентября – 1 октября 2017 года

Редактор М.Ю. Сенашова
Компьютерная верстка: М.Ю. Сенашова

Подписано в печать «15» августа 2017 г. Формат 60 × 90/8. (А4)
Бумага офсетная. Печать плоская.
Усл. печ. л. 11,0. Уч.-изд. л. 11,3.
Тираж 500 экз. Заказ 15

Отпечатано в типографии ИВМ СО РАН
660036, Красноярск, Академгородок