Самоорганизация нейросетей с активными нейронами для обнаружения зависимости активности химических соединений на основе алгоритма поиска аналогов в экспериментальных данных.

А. Г. Ивахненко[1], И. В. Тетко[2] , В.В. Ковалишин[2] , А. И. Луйк[2] ,
Г. А. Ивахненко[3], Н. А. Ивахненко[1]

[1]Кибернетический Центр НАН Украины, Украина,
[2]Институт биоорганической химии и нефтехимии НАН Украины, Украина, 252660 Киев-94,
ул. Мурманская, email: vkov@bioorganic.kiev.ua
[3]Национальный Институт Стратегических Исследований

Абстракт

Разработана структура нейросети с активными нейронами, решающая задачу повышения точности прогноза активности химических соединений. Для прогноза нейросеть использует метод поиска аналогов в экспериментальных данных, причем существенную роль имеет оптимальный выбор множества признаков. Метод оптимизации числа признаков - перебор значений критерия для различных вариантов множества признаков. При малом числе признаков применяется полный перебор всевозможных множеств, а при значительном числе - метод постепенного уменьшения множества признаков. В нейросети осуществлен процесс самоорганизации структуры и выбора параметров.

Введение

Известно, что рациональный поиск соединений, обладающих заданной направленностью биологического действия, требует наличия сведений о связи молекулярной структуры с биологической активностью. Знание этой связи также необходимо для конструирования новых лекарств, поскольку одна из основных предпосылок методов конструирования лекарств - положение о том, что соединения сходной структуры имеют сходные типы биологической активности.

В последние годы резко возрос интерес к рациональным методам исследования связи между структурой и активностью (метод Ханша, квантовохимические методы, и др.). Наряду с этим получили дальнейшее развитие вычислительные методы систематизации и обработки химических данных.

Одним из методов классификации соединений, позволяющие сделать некоторые общие заключения о характере действия новых соединений, стали нейронные сети (НС).[1,2,3] Эти методы позволяют по сравнению с традиционными методами, такими как множественный регрессионный анализ, линейный дискриминантный анализ и др., лучше установить корреляции между структурой и активностью больших групп соединений, структуры которых сильно отличаются. Хотя существует большое количество НС различных топологий, все же в основном большинство работ в этой области посвящено использованию НС с жестко установленной структурой и связями между нейронами (НС с обратным распространением ошибки и др.)[4,5]. Такие типы нейросетей можно назвать субъективными.

В отличие от этого, нейроны, выбирающие свои исходные переменные в результате обучения или самоорганизации можно назвать активными нейронами. Выбор связей активными нейронами однозначно определяет структуру соединений всей нейронной сети[6,7,8]. Нейросеть с активными нейронами представляет собой комитет алгоритмов, каждый из которых решает с некоторой точностью ту же проблему. В каждом ряду такой нейронной сети используются несколько модулей, называемых нейронами. Каждый из

нейронов сети представляет собой элементарную систему, решающую ту же проблему. Нейроны такой сети поставлены в различные условия. Они могут отличаться как по выходным переменным так и множеству входных переменных. Вычислительные модули объединяются в многорядную структуру с целью повысить точность решения задачи за счет более полного использования поступающей на вход информации.[6]

Применение такого типа нейросетей особенно актуально к выборкам зашумленных данных. Для прогноза переменных с различной дисперсией помех требуется построение различного числа рядов активных нейронов сети. Чем точнее измерена переменная, тем меньше дисперсия помех - тем меньше требуется рядов нейросети для достижения максимальной точности прогноза[9].

Шум очень часто присутствует в любых экспериментальных данных, поэтому мы проводили наше исследование химических данных с помощью НС с активными нейронами. В данной работе рассмотрено применение нейросети с активными нейронами для решения задачи повышения точности прогноза активности для двух типов химических соединений.

При самоорганизации каждого из входных нейронов неросети использовался алгоритм поиска аналогов.[10,11] Данный алгоритм относится к алгоритмам безмодельного типа. Роль модели в данном случае выполняет сам исследуемый объект. При этом в случае малой дисперсии помех объект действует как полная физическая модель, а при больших помехах - как упрощенная нефизическая модель. Поэтому прогноз по аналогам можно рекомендовать при любой дисперсии помех - малой или большой.[11]

Перед самоорганизацией нейросети использовалась процедура выбора оптимального множества признаков. В результате количество нейронов сети значительно уменьшалось, поскольку для анализа использовалось только оптимальное количество признаков.

Описание структура нейросети с активными нейронами

В нейросети с активными нейронами отдельные независимые алгоритмы прогноза применяются к входным переменным или признакам, одновременно или последовательно во времени. Понятие выходной переменной, свойственное обычным алгоритмам прогноза, здесь распространяется на все переменные. На Рис. 1 показано, что на всех рядах нейросети используются подобные друг другу нейроны, действующие по алгоритму поиска аналогов. Каждый из нейронов имеет блок конвергенции (на рис. обозначено цифрами 2,5), в котором происходит отбор оптимального множества признаков для каждой переменной и блок прогноза (цифры 3, 6) данной переменной по аналогу. На первый ряд нейросети поступает только исходная выборка данных, тогда как на второй ряд исходная выборка и ее первые прогнозы по каждой переменной. Аналогично на третий ряд подается исходная выборка данных и прогнозы предыдущих рядов нейронов. За каждой конвергенцией числа признаков (блоки 2,3) следует дивергенция числа признаков (блоки 4,7) , дающая возможность на каждом ряду выбрать самые эффективные признаки. Наращивание рядов нейронов продолжается до тех пор, пока точность прогноза интерисующих нас переменных возрастает. Это происходит за счет использования так называемых вторичных факторов, которые обычно представляют собой функции первичных входных факторов. Было установлено, что выходные переменные прогнозирующих алгоритмов также могут быть эффективными факторами, которые можно использовать в последующем ряду нейронов многорядной нейросети.

Алгоритм поиска аналогов

Входная информация представляется в виде выборки исходных данных описывающих ряд химических соединений. Предполагается, что новая исследуемая молекула будет сравниваться со всеми эталонами с тем, чтобы найти наиболее близкий

эталон, который называется аналогом. Химическая активность исследуемой молекулы полагается равной активности эталона.

В многомерном пространстве признаков каждой молекуле соответствует точка. Мерой близости молекул принимается расстояние между их характеристическими точками. Точку, соответствующую координатам указанным в выборке, будем называть характеристической точкой.

Расстояние L от характеристической точки каждой молекулы до характеристических точек остальных молекул определяется как

$$L = \sqrt{\sum_{j=1}^{n} \left(x_{ij} - x_{aj}\right)^2} \qquad (1)$$

где x_{ij} - это величина j признака i анализируемой молекулы, x_{aj} - величина ее аналога, n - количество анализируемых признаков. Аналогом служит молекула, характеристическая точка которой расположена ближе других к точке исследуемой молекулы.

Все молекулы, заданные в выборке экспериментальных данных по очереди считаются исследуемыми молекулами. Для каждой молекулы находится ее аналог (ближайший сосед).

В качестве критерия оценки качества полученной модели использовался *критерий вариации ошибки прогноза* (δ)[12]

$$\delta = \sqrt{\sum_{n=1}^{N} (K_i - K_a)^2 \Big/ \sum_{n=1}^{N} (K_i - \overline{K})^2} \qquad (2)$$

где K_i - степень активности i молекулы, K_a - степень активности аналога, $\overline{K}$ □-среднее значение степени активности всех молекул, N - количество исследуемых молекул. Ясно, что в этом случае активность подлежит измерению и должна быть указана в выборке исходных данных для каждого соединения. Этот критерий позволяет оценить успех апроксимации или прогноза.[12] Если $\delta<0.5$ -результаты моделирования хорошие; если $0.5<\delta<0.8$ результаты моделирования удовлетворительные; при $\delta>1.0$ -моделирование не получилось, поскольку модель приносит дезинформацию.

Метод комплексирования двух аналогов

Комплексированием называется учет аналогов с некоторыми коэффициентами веса, сумма которых равна единице. Если мы через K_0 обозначим активность исследуемой молекулы, через K_1 активность ближайшей к ней, а через K_2 -более удаленную то

$$K_m = a*K_1 + b*K_2 \qquad (3)$$

где $a+b=1$, $a=L_2/(L_1+L_2)$, $b=L_1/(L_1+L_2)$, L_1-расстояние от характеристической точки исследуемой молекулы до точки первого ближайшего аналога с активностью K_1, L_2-расстояние до следующей точки с активностью K_2 , K_m - прогноз молекулы K_0.

Поиск оптимального множества признаков

Обычно при предварительном анализе данных учитываются все признаки, в том числе и те которые, могут не иметь отношения к рассматриваемой классификационной задаче. Поскольку не все признаки существенны для решения рассматриваемой задачи, необходимо найти метод уменьшения их количества. Для решения данной задачи поиск оптимального

множества признаков осуществляется перед самоорганизацией нейросети с целью сокращения количества входных нейронов. Аналогично блоки конвергенции каждого из нейронов осуществляют также поиск оптимального количества признаков из исходной выборки данных и их прогнозов.

Алгоритм последовательно реализует полный перебор всех множеств признаков с выбором наименьшей величины δ на каждом шаге перебора. Однако полный перебор всех возможных множеств признаков можно рекомендовать только при количестве признаков не более 10-12, поскольку в противном случае количество возможных комбинаций слишком велико.

Сокращение объема вычислений достигается при помощи применения итерационной процедуры. С этой целью были рассмотрены два метода перебора.

Первый метод представляет собой процесс постепенного расширения исходного пространства признаков путем постепенного наращивания множества признаков. Если через *M* обозначить число исследуемых признаков, то сначала перебору подлежат *M* одномерных выборок. По минимуму критерия вариации ошибки прогноза δ отбирается наиболее эффективный признак. На следующем шаге перебору подлежат *M-1* двухмерных выборок, включающим в выборку признак полученный для одномерного перебора. Затем производится перебор *M-2* трехмерных множеств, содержащих лучшие признаки из двухмерного перебора и т.д. На каждом шаге находится минимум критерия вариации ошибки прогноза δ. Описанная выше процедура выполняется до тех пор, пока не будет найдено минимальное значение δ.

Второй метод представляет собой процедуру уменьшения множества признаков. В качестве исходного множества признаков принимается полное множество *M*. На каждом шаге происходит уменьшения множества на один признак путем исключения наиболее неэффективного признака. Описанная выше процедура повторяется до тех пор, пока критерий вариации ошибки δ не достигнет минимума что соответствует наиболее оптимальному множеству признаков.

Результаты исследования показывают что алгоритм сужения пространства признаков позволяет найти наиболее оптимальное количество признаков близкое к полному перебору. В качестве примера на *Рис. 2* показаны кривые найденные двумя методами перебора для аналогов антимицина. Кривая *1* получена в результате постепенного увеличения множества признаков, а кривая *2* методом постепенного уменьшения множества. Минимум кривой *2* показывает что наиболее оптимальное количество признаков получено с помощью метода постепенного уменьшения множества признаков.

Результаты исследования

Два набора данных были использованы для анализа. В первом наборе для анализа был использован ряд молекул аналогов антимицина, обладающих ‘antifilaril’ активностью. Исходная выборка данных содержала 31 соединение и 53 признака.[13]

На первом этапе анализа был произведен поиск оптимального числа признаков. В результате оптимизации было найдено множество признаков, содержащее только 9 признаков из 53. На *Рис.2* показано, что оптимальное число признаков получено при постепенном уменьшении множества признаков (кривая *2*) Наименьшая ошибка определения активности молекул δ=0.627 получена для следующего массива признаков {4,6,12,13,19,20,31,33,41}. Найденные множества признаков не только самые простые, но и самые эффективные, т. к. для них достигается наименьшее среднее расстояние между характеристическими точками исследуемого объекта, которыми служат все точки выборки данных по очереди и их первым аналогом.

Для дальнейшего повышения точности прогноза была построена нейросеть с активными нейронами. На каждом ряду нейросети использовалось по десять нейронов. В

Таблице 1 представлены результаты прогноза по каждому из 9 отобранных признаков и активности. Сравнивая цифры по рядам можно сделать вывод, что для прогноза активности необходимо построить не менее 10 рядов нейросети, при этом наименьшая ошибка определения активности молекул δ=0.391. Выделенные курсивом значения показывают какое количество слоев нейросети необходимо построить для повышения точности прогноза каждого признака. Так например признак №33 наименее зашумлен и всего три слоя нейросети необходимо построить. Тогда как минимальное значение δ для признака 12 достигается только на 10 слое нейросети. Поскольку минимальное значение величины критерия вариации ошибки прогноза для активности $\delta<0.5$,[12] то можно сделать вывод о том что результаты моделирования являются хорошими.

Вторая выборка данных - ряд молекул монозамещенных бензолов[14] - состояла из 35 соединений и всего из 31 признака.

Используя алгоритм постепенного сокращения числа признаков было найдено 11 признаков {2-5,11,13,18,20,21,26,30} для которых наименьшая ошибка определения активности молекул равна 0.299. (Рис. 3).

В Таблице 2 приведены сравнительные результаты прогноза при построении нейросети для всех признаков (набор №1) и множества признаков полученного с помощью алгоритма оптимизации (набор №2).

На втором этапе анализа использовалась выборка данных, состоящая из 11 признаков, найденная Давидом Ливинстоуном и др. исследователями из исходной выборки данных (см. детали в **Ошибка! Закладка не определена.**). Результаты анализа с помощью нейросети приведены в Таблице 2 (набор №3). С помощью метода полного перебора признаки 5 и 9 были исключены из начальной выборки данных. (набор №4). Для нахождения минимума критерия δ=0.211 для этого набора необходимо было построить нейросеть с 10 слоями нейронов. Незначительное снижение δ (от 0.299 до 0.211) свидетельствует о незначительной зашумленности исходных данных.

Полученные результаты показывают что нейросети построенные для полного числа признаков и числа признаков найденного с помощью алгоритма оптимизации практически одинаковы (см. Таблицу 2). Однако использование алгоритма оптимизации позволяет снизить время расчетов в десятки раз, что является очень существенным при анализе выборок данных очень большой мерности. При значительной зашумленности исходной выборки данных необходимо значительно больше слоев нейросети для фильтрации значений всех признаков. Например, для аналогов антимицина снижение δ от 0.627 до 0.391 потребовало не менее 10 слоев нейронав, тогда как для монозамещенных бензолов только 4 слоя.

На основе выше приведенных результатов можно заключить что использование алгоритма поиска аналогов может быть успешно использовано для повышения точности прогноза активности новых соединений с помощью нейросетей с активными нейронами.

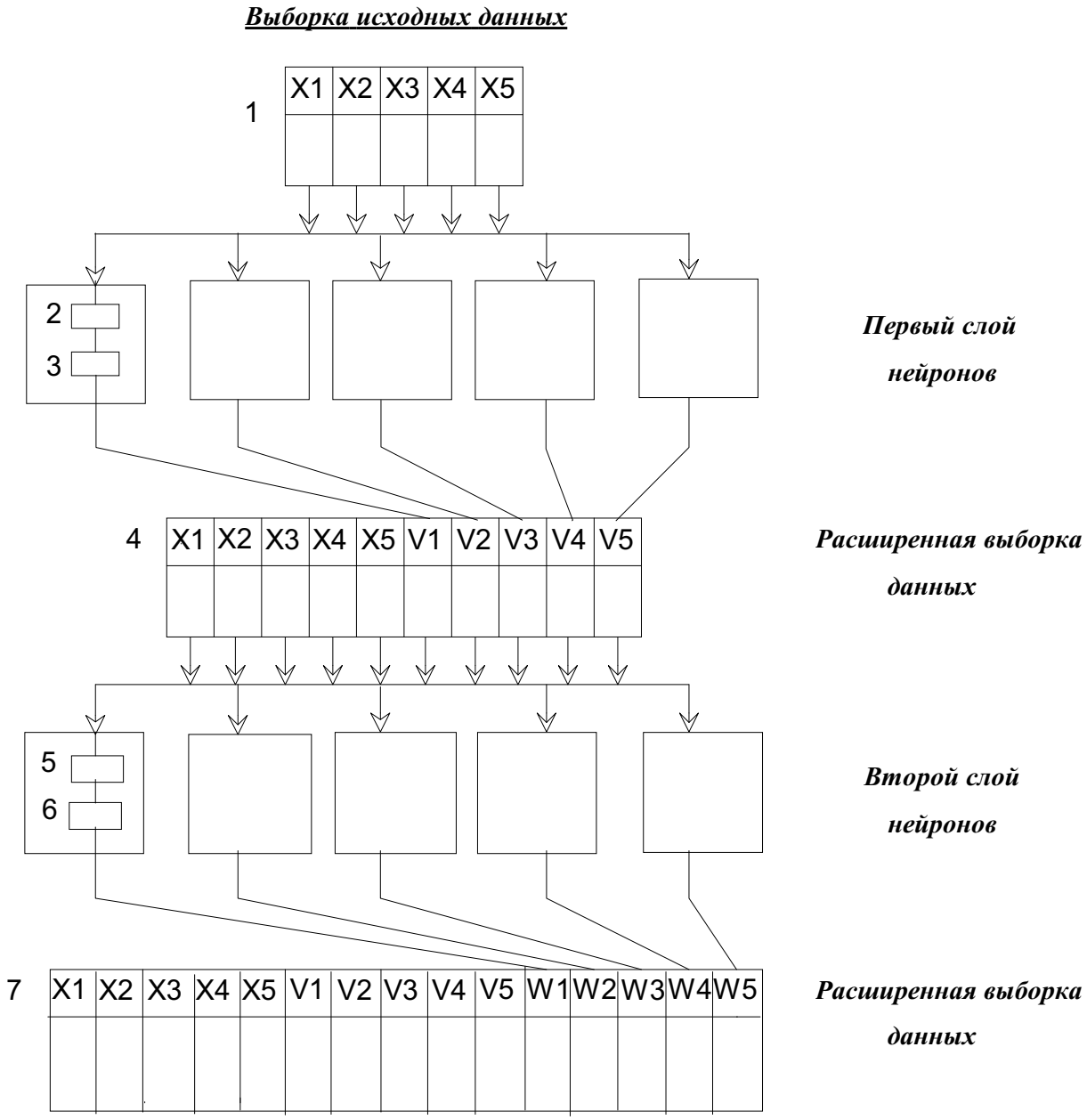

Рис. 1. Схема первых двух слоев нейросети с активными нейронами, действующими по алгоритму поиска аналогов**.**

1-выборка исходных экспериментальных данных; 2-блок первой конвергенции числа признаков, сокращающие множество входных переменных до оптимального размера для каждой переменной; 3 -блок прогноза каждой переменной по аналогу; 4- выборка, содержащая исходные данные и их первые прогнозы; 5-блок конвергенции для второго ряда нейронов; 6- блок прогноза по аналогу;7- выборка, содержащая исходные данные и их прогнозы на первом и втором ряду.

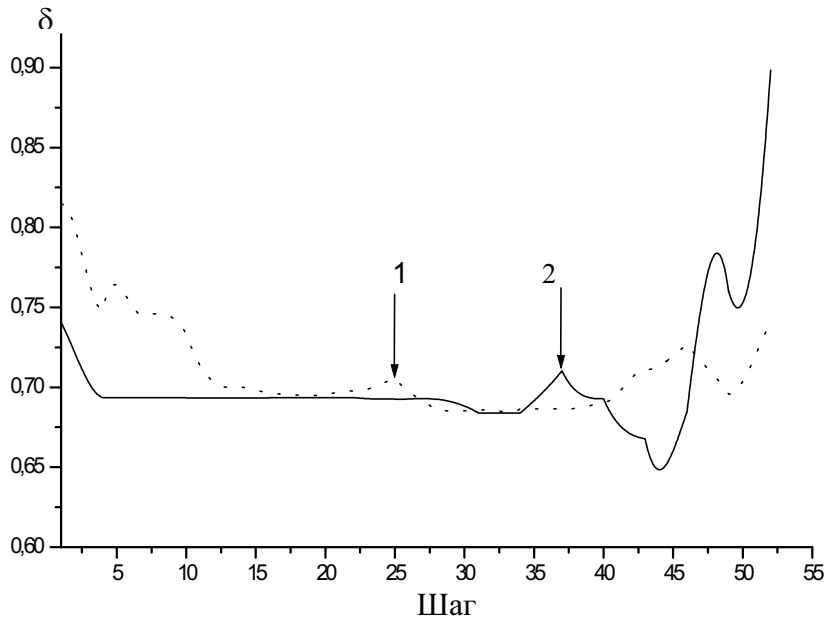

Рис 2. Зависимость критерия вариации ошибки прогноза δ от количества учитываемых признаков для аналогов антимицина**;** 1-кривая получена при отборе методом постепенно усложняющихся множеств; 2-.кривая получена методом постепенного уменьшения множества признаков.

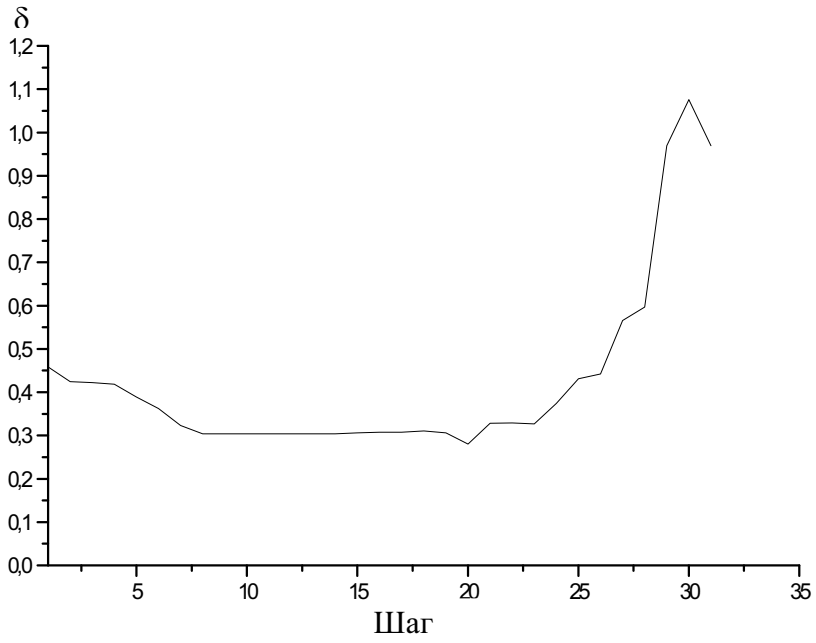

Рис. 3 Зависимость критерия вариации ошибки прогноза δ от количества учитываемых признаков при отборе методом постепенно уменьшения множества признаков для монозамещенные бензолы;

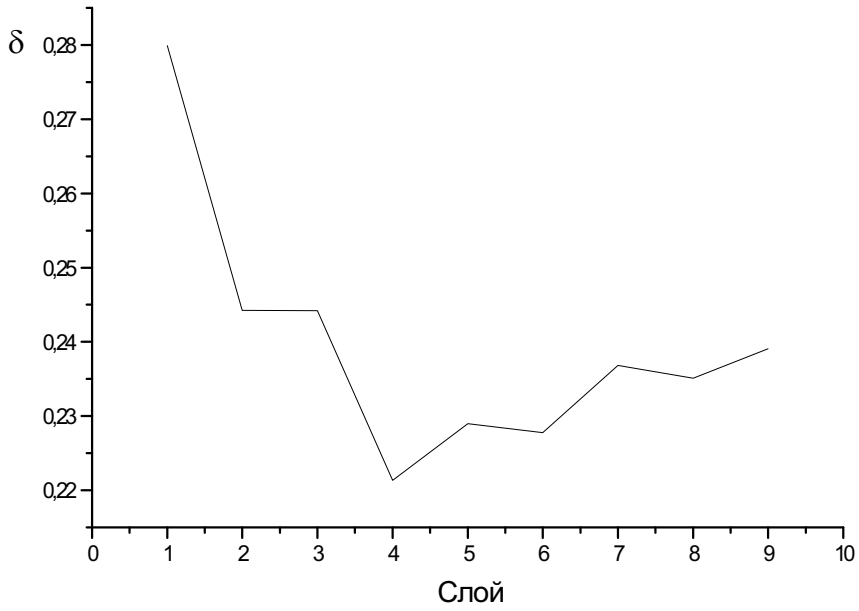

Рис 4. Зависимость критерия ошибки δ от количества слоев нейросети для набора из 11 признаков для монозамещенных бензолов.

Слой	Номер признака									
нейр.	4	6	12	13	19	20	31	33	41	Акт.
1	0.665	0.299	0.632	1.002	0.683	0.697	0.274	0.932	0.929	0.627
2	0.422	0.252	0.681	0.963	0.591	0.488	0.209	0.903	0.942	0.600
3	0.328	0.249	0.506	0.888	0.503	0.474	0.208	**0.901**	0.942	0.627
4	0.311	0.209	0.478	0.888	0.438	0.472	0.207	0.901	0.917	0.483
5	0.305	0.182	0.445	0.851	0.455	0.465	0.206	0.902	0.916	0.482
6	**0.303**[a]	0.176	0.395	0.832	0.433	0.466	0.211	0.901	0.872	0.423
7	0.307	**0.173**	0.364	**0.832**	**0.419**	0.583	**0.203**	0.902	**0.857**	0.421
8	0.306	0.173	0.397	0.832	0.419	**0.458**	0.206	0.902	0.877	0.421
9	0.306	0.173	0.357	0.855	0.429	0.458	0.203	0.902	0.877	0.412
10	0.305	0.173	**0.324**	0.855	0.435	0.469	0.203	0.902	0.877	**0.391**

Таблица 1. Величина критерия вариации ошибки δ для каждого из признаков на каждом слое нейросети.[a] Жирным шрифтом выделено минимальное значение ошибки для каждого признака.

№	Набор признаков		крит. вар.	Количество
	состав	количество	ошибки δ	слоев
1	все признаки	31	0.212	4
2	2-5,11,13,18,20,21,26,30	11	0.221	4
3	все признаки	11	0.225	9
4	1,2,3,4,6,7,8,10,11	9	0.211	10

Таблица 2. Результаты прогноза биологической активности монозамещенных бензолов для 2 наборов данных.

[1] Maddalena, D. (1996). Applications of Artificial Neural Networks to Quantitative Structure Activity Relationships. *Exp. Opin. Ther. Patents.,* 6, 239-251.

[2] Munk M. E. (1996). The Neural Network as a Tool for Multispectral Interpretation. *J. Chem Inf. Comput. Sci*., 36, 231-238.

[3] Fahlman, S. E.; Lebiere, C. (1990). The Cascade-Correlation Learning Architecture.*NIPS*2* Touretzky, D. S., Ed.; Morgan-Kaufmann: San Mateo, CA, pp 524-532.

[4] Tetko, I. V., Villa, A. E. P., & Livingstone, D. J.(1996).; Neural Network Studies. 2. Variable Selection. *J. Chem. Inf. Comput. Sci.,* 36,794-803.

[5] Tollenaere, T. SuperSAB: (1990). Fast Adaptive Back Propagation with Good Scaling Properties. *Neural Networks.* 3, 561-573.

[6] Ivakhnenko, A. G., Ivaxnenko, G.A., Muller, J.-A. (1994). Self-Organization of Neural Networks with Active Neurons. *Pattern Recognition and Image Analysis*, Vol. 4, № 2, pp. 185-196

[7] Lemke, F. (1998). Application of self-Organizing Modeling for Portfolio Prediction. (in press, SAMS).

[8] Ivakxnenko, A. G., Ivakxnenko, G. A.(1996). Normative forecast and optimal multi criteria control using selforganization of systems of nonphysical models. *Problems of control and information*, 28(1-2),2735.

[9] Ivaxnenko, G. A. (1995). Self-organization of neuronet with active neurons for effects of nuclear test explosions forecastings. *System Analysis modeling Simulation (SAMS),* 20, pp. 107-116.

[10] Ivaxnenko, A. G.(1991). An Inductive Sorting Method for the Forecasting of Multidimential Random Processes and Events with the Help of Analog Forecast Complexing. *Pat. Rec. Image Anal.,* Vol.1, №1, pp. 99-108.

[11] Ивахненко, А. Г., Богаченко, Н. Н., Мин, Л. Т. (1997). Этапы оптимизации алгоритма прогнозирования случайных процессов с помощью комплексирования аналогов. *Проблемы управления и информатики*, №4, стр 111-118.

[12] Belogurov, V.P.(1990). Criterion of Model Suitability for Forecasting Quantitative Processes. *Soviet J. Autom. Inform. Sci., V*ol. 24,3, pp. 21-25.

[13] Selwood, D. L., Livingstone, D. J., Comley, J. C. W., O'Dowd, A. B., Hudson, A. T., Jackson, P., Jandu, K. S., Rose, V. S., & Stables, J. N. (1990). Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem*. , 33, 136-142.

[14] Livingstone, D. J., Evans D. A., Saunders, M. R.(1992). Investigation of a Charge-transfer Substituent Constant Using Computational Chemistry and Pattern Recognition Techniques. *J. Chem. Soc. Perkin Trans.,*2, pp. 1545-1550.