Ивахненко Г.А.

Алгоритм комплексирования аналогов для самоорганизации дважды многорядных нейронных сетей

Алгоритм МГУА с комплексированием аналогов используется для прогноза, экстраполяции и распознавания образов плохо определенных объектов. В случае недостаточной априорной информации, неточных измерений, зашумленных и коротких выборок данных лучшие результаты достигаются при помощи поиска аналогов в предыстории и нахождением нефизических моделей. Рассматривается пример нахождения оптимального множества переменных для экстраполяции химического взаимодействия.

Многие объекты распознавания и управления в экономике, экологии, биологии и медицине недетерминированы или нечетки. Они могут быть представлены детерминированной (робастной) частью и дополнительными черными ящиками, действующие на каждом выходе объекта. Единственная информация об этих ящиках - то, что они имеют ограниченные значения выходных переменных, подобные соответствующим состояниям объекта. Согласно [2] разнообразие системы управления или модели должно быть не меньшим, чем разнообразие самого объекта. Закон адекватности, данный С.Биром [1], устанавливает, что для оптимального управления объекты должны быть компенсированы соответствующими черные ящиками системы управления. Для оптимального распознавания образов и кластеризации необходима только частичная компенсация. Чаще всего исследователи стараются минимизировать степень компенсации различными средствами для повышения точности результатов. Равная нечеткость модели и объекта достигается автоматически, в случае если сам объект используется для прогноза. Это достигается с помощью поиска аналогов соответствующими физической модели в выборке данных. В этом случае прогнозы не рассчитываются обычным способом, а находятся в выборке наблюдений. Главные предположения при этом следующие:

- Объект исследования описывается многомерным процессом;
- Многомерный процесс достаточно представителен, то есть основные системные переменные включены в выборку данных и она содержит достаточно много наблюдений;
- Возможно повторение части прошлого поведения системы в будущем.

Алгоритм комплексирования аналогов рекомендуется, когда количество входных переменных больше чем количество наблюдений, в другом случае могут использоваться также параметрические алгоритмы МГУА [3]. Если возможно найти одну или несколько аналогичных частей в прошлом (“аналогичные образы”) для последней части траектории поведения (“начального образа”), то тогда может быть найден прогноз, используя известные продолжения предыдущих аналогичных образов (Рис. 1).

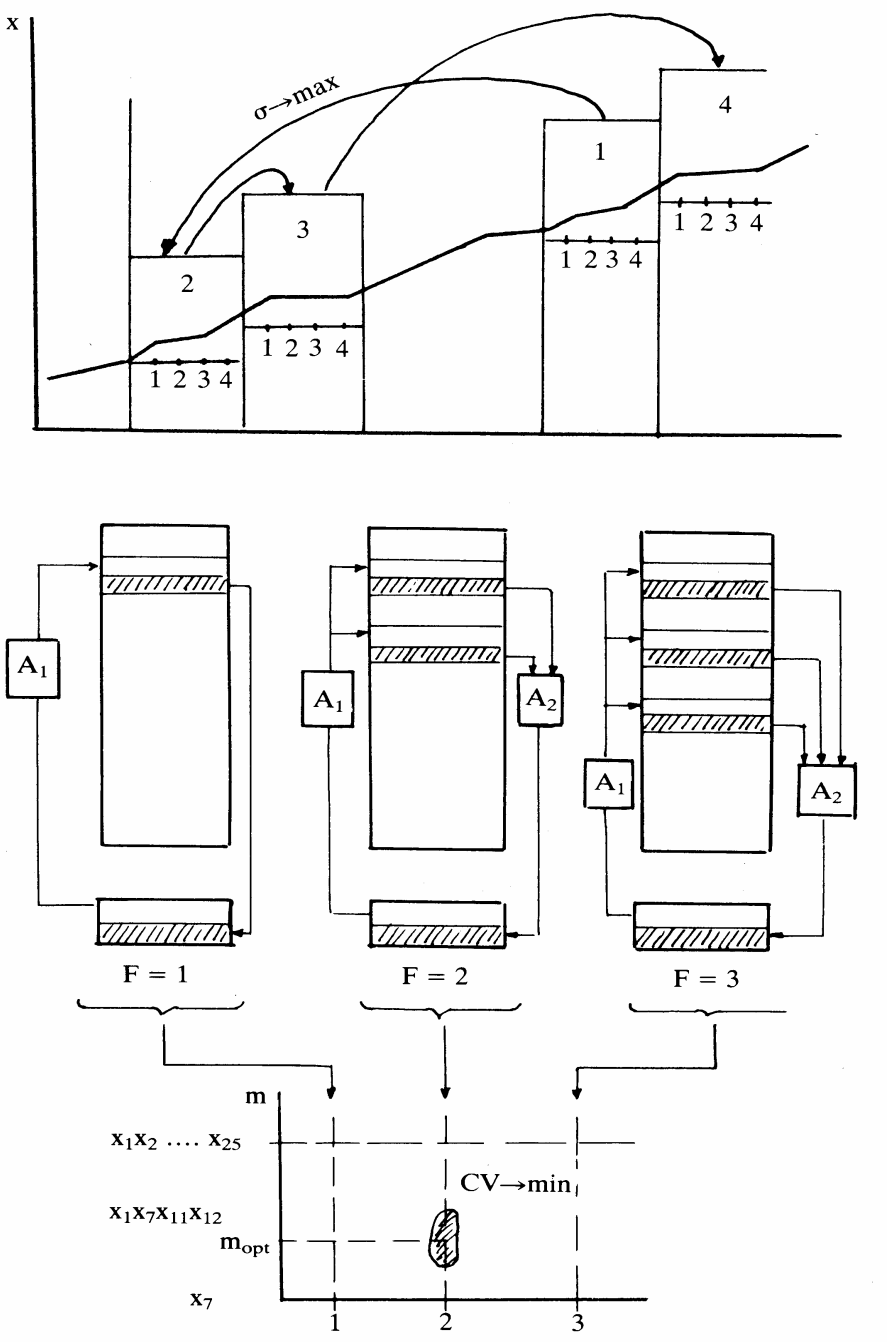

Рис. 1. Схема алгоритма комплексирования аналогов.

Образ A_1, самый близкий к заданному выходному образу A_0, называется его первым аналогом. Следующий по расстоянию образ A_2 называется вторым аналогом и так далее. Образ, которая следует за первым аналогом A_1 по времени, A_{1F}, называется первым прогнозом аналога. Образ, которая следует за вторым аналогом A_2 по времени A_{2F}, называется вторым прогнозом аналога и так далее. Прогноз рассчитывается при помощи комплексирования оптимального количества прогнозов аналога. Используя скользящее окно, которое генерирует множество возможных образов $\{A_{i,\ k+1}\}$, где $A_{i,k+1}= (x_i,x_{i+1},x_{i+2},\ldots,x_{i+k})$ и k+1 - ширина скользящего окна i-го образа, получаем выходной образ $A_k^A = A_{N-k,k+1}$.

Алгоритм выбора аналогичного образа применяется со следующей целью: для данного выходного образа A_k^A необходимо выбрать наиболее подобные образы $A_{i,k+1}, i \in J$ и с их помощью оценить точность прогноза или экстраполяции.

Дважды многорядная нейросеть значительно повышает точность решения задач моделирования. Она используется для оптимизации параметров алгоритма комплексирования аналогов и выбора оптимального множества переменных.

Общая задача оптимизации характеристик алгоритма комплексирования аналогов должна решатся в четырехмерном пространстве перебора. Требуется найти оптимальные, по критерию ошибки, значения следующих четырех групп характеристик:

1. числа и состава множества учитываемых переменных (факторов);
2. числа моментов времени, используемых при расчете аналогов;
3. числа комплексируемых аналогов;
4. значений коэффициентов веса комплексирования прогнозов.

Однако рассмотрение примеров показало, что взаимные связи указанных характеристик таковы, что большой четырехмерный перебор можно заменить двумя одномерными и одним двумерным переборами.

Одномерный перебор множеств входных переменных (факторов) служит наиболее простым средством оптимизации алгоритма комплексирования прогнозов. Этот перебор выполняется для одного первого аналога, при учете одной строки выборки (одного момента) и при заданных значениях коэффициентов веса. При этом критерий рассчитывается по всей длине выборки. Таким образом решается проблема нахождения кратковременного прогноза на один шаг вперед. Более трудной является задача долгосрочного пошагового прогноза случайных процессов. Разработаны следующие процедуры для выбора подобных образов из всех возможных образов во временных рядах.

А. Поиск оптимального множества переменных

Выбор оптимального множества переменных может быть выполнен с помощью перебора возможных множеств переменных. В случае малого числа переменных (<10) используется полный перебор возможных множеств переменных, а для большего количества переменных - используется метод постепенного усложнения множеств переменных (усеченный перебор).

Сначала находится расстояние L от каждого наблюдения до других наблюдений. Евклидово расстояние определяется как:

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

где x_{ij} - значение j-й переменной в i-й анализируемой точке; x_{aj} - значение его аналога; n - количество наблюдений.

В качестве аналога используется наблюдение, которое расположено наиболее близко к данному. Для всех наблюдений в выборке данных находятся их аналоги (ближайшие соседи). В качестве аналога критерия для выбора оптимального множества переменных используется значение критерия точности:

$$RR(s) = \sqrt{\frac{\sum_1^N \left(x_i - x_a\right)^2}{\sum_1^N \left(x_i - \bar{x}\right)^2}} \to \min, \qquad (2)$$

где: x_i - текущее наблюдение; x_a - соответствующий аналог; $\bar{x}$ - среднее значение.

Сокращение процедуры перебора достигается постепенным усложнением множеств переменных, используя переменные, отобранные на предыдущем шаге. Если определить M как число переменных, то на первом ряду будет перебрано M выборок. Тогда M-1 множеств всех пар переменных с переменной, отобранной при предыдущем шаге, используются для поиска аналогов. Затем проверяются N-2 множеств, которые включают лучшие переменные из предыдущих рядов и так далее. Результаты вычисления на примере показали, что оба типа перебора дали практически тот же самый результат (таблица 1). В таблице рассматриваются результаты полного и усеченного перебора для реальных химических данных. Обе процедуры выбирают по 9 переменных, исключая переменные x_5 и x_9.

Таблица 1. Значения критерия точности RR(s) при полном и усеченном переборе

	Полный перебор признаков	
1	6	0.3194
2	1, 2	0.1687
3	3, 6, 8	0.1915
4	1, 2, 3, 4	0.1751
5	1, 2, 3, 7, 8	0.1795
6	1, 2, 3, 5, 6, 10	0.1691
7	1, 2, 3, 4, 7, 10, 11	0.1466
8	1, 2, 3, 4, 7, 8, 10, 11	0.1471
9	1, 2, 3, 4, 6, 7, 8, 10, 11	**0.1416**
10	1, 2, 3, 4, 5, 6, 8, 9, 10, 11	0.1930
	Ограниченный перебор признаков	
1	6	0.3194
2	3, 6	0.2662
3	3, 6, 8	0.1915
4	3, 6, 8, 11	0.2178
5	3, 6, 8, 10, 11	0.2477
6	1, 3, 6, 8, 10, 11	0.2373
7	1, 2, 3, 6, 8, 10, 11	0.1981
8	1, 2, 3, 6, 7, 8, 10, 11	0.1660
9	1, 2, 3, 4, 6, 7, 8, 10, 11	**0.1416**[a]
10	1, 2, 3, 4, 6, 7, 8, 9, 10, 11	0.1975
11	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11	0.2014

B. Преобразование аналогов

Этот шаг не является обязательным. Большинство реальных процессов в больших системах имеют эволюционный характер. В этом случае не выполняется стационарность, важное условие успешного использования метода комплексирования аналогов. Временные ряды могут состоять из нестационарных образов с подобными формами, но иметь различные средние значения, стандартные отклонения и тренды. В литературе рекомендуется оценивать зависимость разности между процессом и его трендом, которая является неизвестной функцией времени. В этом случае результаты метода комплексирования аналогов зависят от вида выбранной функции тренда.

При этом желательно определить преобразованные образы $A^*_{i,k+1} = (\underline{x}^*_i, \underline{x}^*_{i+1}, ... \underline{x}^*_{i+k})$, где $x^*_{jt} = w_{j0} + w_{j1} x_{jt}$. Веса w_0, w_1 для каждого образа $A_{i,k+1}$, когда k>1, могут быть оценены посредством МНК, который оценивает не только неизвестные веса, но также и сумму квадратов s_i^2, которая может быть использована на следующем шаге как мера подобия.

C. Выбор наиболее подобных аналогов

Самый близкий аналог называется первым аналогом A_1, следующий по расстоянию A_2 называется вторым аналогом и так далее до последнего аналога A_F. Расстояния могут быть измерены с помощью евклидовой меры близости между выходным образом и его аналогом.

В нашем случае не требуется находить меру близости между аналогами, так как полная сумма квадратов s_i^2 дает нам информацию о близости между A_1 и A_F.

D. Объединение прогнозов (комплексирование)

Каждый отобранный аналог имеет свое продолжение во времени, по которому строится прогноз. Таким способом мы получаем F прогнозов, необходимых для комлексирования. В литературе есть несколько способов объединения прогнозов. Неизвестные прогнозы x_{N+k} для M переменных могут быть найдены как линейная комбинация продолжений выбранных аналогичных образов, то есть:

$$x_{N+j} = g_0 + \sum_{j \in J} g_k x^*_{j+k+i},$$

Неизвестные параметры g_0, g_j, $j \in J$, оцениваются с помощью параметрических процедур выбора, или используя индуктивные алгоритмы самоорганизации. Единственная проблема может быть в малом числе наблюдениях или малом числе отобранных образов.

2. Самоорганизация дважды многорядных нейронных сетей

Дважды многорядные нейронные сети предназначены для решения различных задачи. Это может быть идентификация зависимостей (аппроксимация), распознавание образов и

ситуаций, или прогноз случайных процессов и повторяющихся событий по информации, содержащейся в выборке данных испытаний или управления объектом.

Современные компьютерные технологии позволили создать новый подход в нейронных сетях, который увеличивает точность классических алгоритмов моделирования. Такие многорядные системы могут решать сложные задачи. Мы можем использовать алгоритмы МГУА как активные нейроны, для которых процессы самоорганизации хорошо изучены. Алгоритмы МГУА - примеры сложных активных нейронов, потому что они сами выбирают эффективные входы и соответствующие им коэффициенты в процессе самоорганизации. Проблема самоорганизации структуры связей нейросети решается довольно простым способом.

Каждый нейрон - элементарная система, которая решает ту же самую задачу. Цель, достигаемая в объединении многих нейронов в сеть, состоит в том, чтобы повысить точность решения данной задачи с помощью лучшего использования входных данных. Как уже отмечалось, функция активных нейронов может быть выполнена различными системами распознавания, например, перцептронами Розенблатта с двумя рядами, решающими задачи распознавания образов. При самоорганизации нейронной сети вначале применяется полный перебор, чтобы определить количество рядов нейронов и множества входных и выходных переменных для каждого нейрона. Минимум критерия выбирает переменные, по которым следует строить нейронную сеть и сколько рядов нейрона следует использовать. Таким образом, самоорганизация нейронной сетей во многих отношениях подобна такой же процедуре для каждого активного нейрона.

Активные нейроны способны в течение процесса самоорганизации оценивать, какие входы необходимы, чтобы минимизировать данную объективную функцию нейрона. В нейросети с такими нейронами мы будем иметь дважды многорядную структуру: сами нейроны многорядные, и они будут объединены во многорядную сеть. Они могут обеспечивать генерацию новых свойств особого типа (выходы нейронов предыдущего ряда) и выбора эффективного множества переменных на каждом ряду нейронов. Выходные переменные предыдущих рядов - очень эффективные вторичные входы для нейронов следующего ряда. Ряды активных нейронов действуют подобным фильтру Калмана: множество выходных переменных повторяет входное множество, но с фильтрацией помех. Количество активных нейронов в каждом ряду равно количеству переменных в исходной выборке данных.

Структура нейросети показана на рис. 2. Выборка расширяется только благодаря включению выходных переменных, вычисленных на каждом предыдущем ряду нейросети. В выборках показана форма дискретного шаблона, используемого для обучения нейронов по алгоритму комплексирования аналогов МГУА. Алгоритм выбирает, какой из предложенных параметров должен быть учтен, и оценивает коэффициенты связи между ними.

Каждый ряд в нейронной сети содержит нейроны, чьи выходы соответствуют каждый определенной переменной: выход первого нейрона для первой переменной, выход второго нейрона для второй переменной и так далее. Каждый столбец состоит из нейронов, чьи выходы соответствуют одной из переменных. В свою очередь, от каждого столбца выбирается один нейрон с минимальным значением критерия вариации. Более детально, из первого столбца нейронов, для которых выходом является первая переменной, выбирается один нейрон, имеющий наибольшую точность; точно так же один нейрон выбирается из второго столбца нейронов, для которых выходом является вторая переменная, и так далее. Эта процедура выбора однозначно определяет количество рядов для каждой переменной и, таким образом, структуру нейронной сети.

Для начала строится первый ряд нейронов в нейросети и определяется, насколько будет точный прогноз для каждой переменной. Для этой цели мы используем дискретный шаблон, который дает запаздывание на один или два шага для всех переменных. Затем мы последовательно добавляем ряды к нейронной сети, как показано на рис.2, и продолжаем это до тех пор, пока улучшается прогноз или уменьшается значение внешнего критерия.

Для каждого нейрона мы применили расширенную процедуру определения одной модели (из пяти наиболее близких к оптимальной модели). Для оптимальных моделей мы вычислили критерий вариации ошибки прогноза $RR(s)$. Можно показать, что нет необходимости строить нейронную сеть для прогнозирования тех переменных, для которых значение критерия вариации имеет наименьшее значения на первом ряду. Лучше строить нейронную сеть для прогнозирования тех переменных, для которых критерий вариации имеет наименьшее значение на последних рядах нейросети.

Уравнения нейронов сети определяют связи, которые должны быть реализованы в нейронной сети; таким образом они помогают решать задачи структурной самоорганизации нейронной сети. Для краткости, выборка данных в упомянутом примере расширена только одним способом: выходные переменные первого ряда передаются как дополнительные переменные второму, третьему и т.д. ряду нейронов. Можно сравнивать различные схемы расширения выборки данных по значению внешнего критерия.

Задача самоорганизации дважды многорядных нейросетей с активными нейронами состоит в том, чтобы оценить число рядов активных нейронов и множества возможных входов и выходов каждого нейрона. Переборная характеристика "количество рядов нейросети – значение внешнего критерия" определяет оптимальное количество рядов отдельно для каждой переменной. Нейросети с активными нейронами могут быть применены для того, чтобы повысить точность краткосрочных и долгосрочных прогнозов.

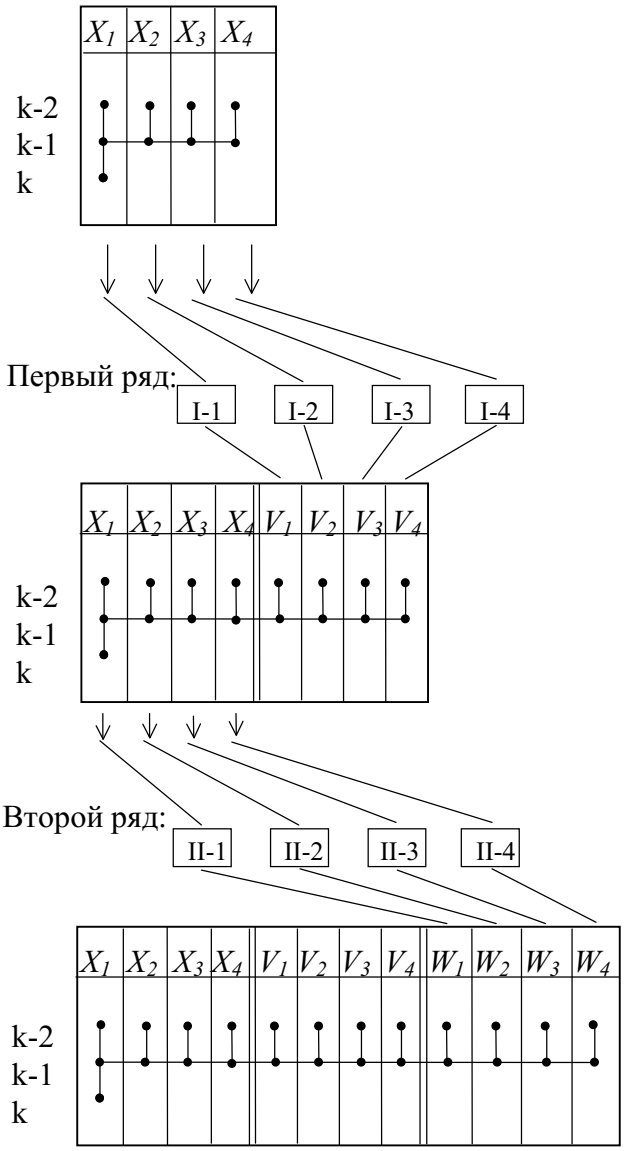

Рис. 2. Схема первых двух рядов дважды многорядной нейронной сети.

Не только алгоритмы МГУА, но и многие другие алгоритмы моделирования и распознавания образов также могут быть использованы как активные нейроны. Их точность может быть увеличена двумя способами:

- Каждый выход алгоритма (активного нейрона) генерирует новую переменную, которая может использоваться как новый признак в следующих рядах нейросети;
- Множество признаков может быть оптимизировано на каждом ряду. Признаки (включая вновь сгенерированные) могут быть ранжированы по их эффективности и несколько из наиболее эффективных признаков могут использоваться как входы для следующих рядов нейронов. В обычной однократно многорядной нейросети множество входных переменных может быть выбрано только один раз.

3. Процедура "расширения и сужения" выборки данных

Количество переменных расширяется на каждом следующем ряду нейросети (от двух до четырех в нашем примере). Вместе с тем переменные, которые являются менее эффективными, должны быть исключены в каждом ряду нейросети (две переменные в примере). Для этой цели все переменные-кандидаты должны быть проранжированы по способности прогнозировать с помощью комплексирования аналогов, и только определенное количество наиболее эффективных переменных (равное четырем в примере) должно быть включено в полное множество переменных каждого следующего ряда нейросети. В обычной нейросети (перцептроне) множество входных переменных оптимизируется только один раз на первом ряду. В дважды многорядных нейросетях с активными нейронами множество переменных может быть оптимизировано несколько раз, на всех рядах нейронов. [4].

4. Пример

Описанный выше метод был применен для поиска оптимального множества переменных для экстраполяции взаимодействия химических соединений на двух выборках данных [5]. В первой выборке для анализа был использован ряд молекул - аналогов антимицина (имеющих действие антифиларила).

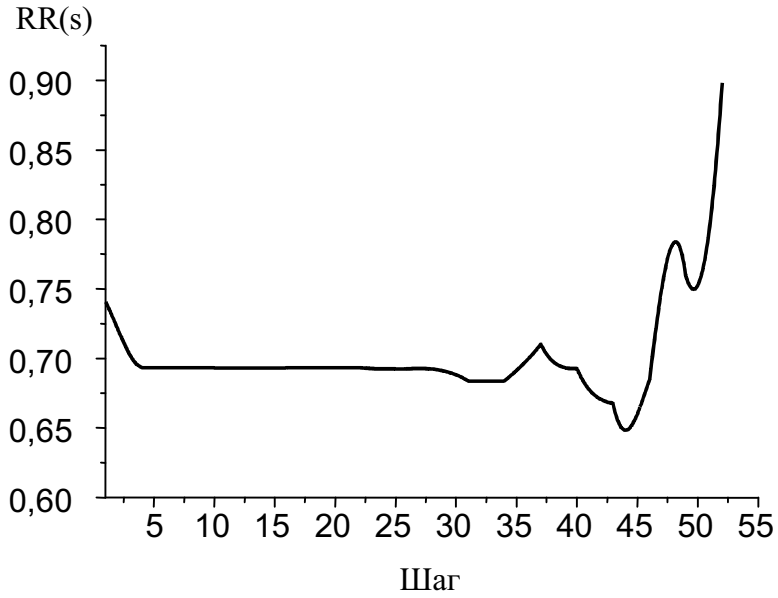

Рис 3. Процедура оптимизации множества входных переменных для аналогов антимицина для первого ряда нейросети с активными нейронами.

Вначале входная выборка данных содержала 31 соединение и 53 переменные. В результате усеченного перебора было найдено подмножество, которое содержит только 10 переменных. На рисунке 3 показаны зависимости $RR(s)$ от шагов перебора. На каждом шаге число рассматриваемых переменных увеличивалось на 1 переменную. На последнем ряду

проверяются все переменные. В рисунке показано, что $RR_{\min}$= 0.647 был получено для 10 переменных X = {3,13,19,23,24,27,29,32,50,52}. Следует отметить, что значение критерия увеличивается при дальнейшем увеличении числа переменных.

Вторая выборка данных состояла из 35 соединений и 31 переменной. Минимальное значение критерия для определения взаимодействия молекул, равное 0.1237, было получено для следующего множества переменных: X = {2,7,9,14,16,20,25,26,30}.

Найденные множества переменных не только наиболее простые, но и наиболее эффективные, потому что для них достигается минимальное среднее расстояние между характерными точками объекта исследования, которое последовательно использует все точки выборки и их первые аналоги.

С помощью вычислительных экспериментов был найдено, что дисперсия шума для выходной переменной на один порядок меньше, чем дисперсия шума во входных данных. Это подтверждает, что аналоги могут использоваться для фильтрации шума [6].

4. Заключение

Эффективность применения алгоритма комплексирования аналогов была многократно подтверждена при решении задач прогнозирования случайных событий и кластеризации вместо обычных алгоритмов математического моделирования [8]. Этот алгоритм может быть применен для прогноза всех переменных в исходной выборке данных с разной эффективностью. Три оптимизации благодаря перебору помогают увеличить точность прогноза. Для дальнейшего увеличения точности может использоваться нейросеть с активными нейронами, где нейроны - алгоритмы комплексирования аналогов. Нейросеть должна рассматриваться как инструмент для повышения точности прогнозирования процессов, благодаря генерации новых переменных (X_{1F}, X_{2F}, $X_{1FF,}$…) и выбору наиболее эффективных множеств переменных в каждом ряду нейросети. С биологической точки зрения, нейросети с аналоговыми активными нейронами более подобны мозгу, чем искусственные нейросети использующие математическое моделирование.

Литература

1. Beer S. Cybernetics and Management, English Univ. Press, London, 1959, p.280.
2. Ashby D. An introduction to cybernetics. J. Wiley, New York 1958.
3. Madala H.R. and Ivakhnenko A.G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press Inc., Boca Raton, 1994.
4. Ivakhnenko A.G. An Inductive Sorting Method for the Forecast of Multidimensional Random Processes and Analog Events with the Method of Analog Forecast Complexing. Pattern Recognition and Image Analysis, 1991, Vol.1,No1.pp.101-107.
5. Ivakhnenko A.G., Kovalishyn V.V., Tetko I.V., Luik A.I., Ivakhnenko G.A., Ivakhnenko N.A. Self-Organization of Neural Networks with Active Neurons for Bioactivity of Chemical Compounds Forecasting by Analogues Complexing GMDH algorithm. Report at the ICANN'99 Conference, London.
6. Ивахненко А.Г., Богаченко Н.Н., Ли Тянь Мин. Безмодельное прогнозирование случайных процессов при помощи комплексирования прогнозов по аналогам. // Проблемы управления и информатики, № 4, 1997, с.111-188.
7. Ivakhnenko A.G. Ivakhnenko G.A. and Mueller J.A. Self-organization of neuronets with Active Neurons. Pattern Recognition and Image Analysis,1994,Vol.4,No.4,pp.177-188.
8. Mueller J.-A., Lemke F. Self-Organizing Data Mining. Libri, Hamburg, 2000, ISBN 3-89811-861-4.
9. Ivakhnenko G.A. Self-organization of Neuronet with Active Neurons for Effects of Nuclear Tests Explosions Forecasting. System Analysis Modeling Simulation SAMS, 1995, Vol.20, No.4, pp.107-116.

Г.А. Ивахненко

Алгоритм комплексирования аналогов для самоорганизации дважды многорядных нейронных сетей

Алгоритм МГУА с комплексированием аналогов используется для прогноза, экстраполяции и распознавания образов плохо определенных объектов. При недостаточной априорной информации, неточных измерениях, зашумленных и коротких выборках данных лучшие результаты достигаются при помощи поиска аналогов в предыстории и построения нефизических моделей. Рассматривается пример нахождения оптимального множества переменных для экстраполяции химического взаимодействия.

Г.О. Івахненко

Алгоритм комплексування аналогів для самоорганізації двічі багаторядних нейронних мереж

Алгоритм МГУА з комплексуванням аналогів використовується для прогнозу, екстраполяції та розпізнавання образів погано визначених об'єктів. У випадку недостатньої апріорної інформації, неточних вимірів, зашумлених і коротких вибірок даних кращі результати досягаються за допомогою пошуку аналогів у передісторії та побудови нефізичних моделей. Розглядається приклад знаходження оптимальної множини змінних для екстраполяції хімічної взаємодії.

IvakhnenkoG.A.

Analogues complexing algorithm for twice-multilayer neural networks self-organization

The Analogues Complexing GMDH algorithm is used for forecasting, extrapolation and pattern recognition of ill-defined objects. In the case of insufficient a priori information, not very accurate measurements, noisy and short data sample, better results are reached by search of analogues in prehistory and by construction of non-physical models using this analogues. The example of optimal set of variables construction for extrapolation of chemical activity of two data samples is considered.