



УДК 681.3:06

**СТРУКТУРИЗАЦІЯ І АНАЛІЗ ДАНИХ В РАСТУЩИХ  
ПИРАМИДАЛЬНИХ СЕТЯХ**

**В.П. ГЛАДУН, Н.Д. ВАЩЕНКО, В.Ю. ВЕЛИЧКО, Ю.Г. ТКАЧЕНКО**

Рассматривается организация памяти интеллектуальных систем в виде растущих пирамидальных сетей (РПС), обеспечивающая формирование понятия, иерархическое упорядочение, классификацию вводимой информации. Приведены алгоритм построения РПС при вводе признаковых описаний объектов по частям, а также результаты применения РПС для анализа данных в химии и материаловедении.

**ВВЕДЕНИЕ**

Соревнование компьютера с человеком в решении так называемых интеллектуальных задач все чаще завершается победой компьютера. Однако возникает очевидное противоречие — компьютер решает сложные многовариантные задачи в искусственных, относительно бедных средах (например, шахматы), и «спотыкается» в решении простых (для человека) житейских задач, требующих, тем не менее, быстрого понимания и оценки многокомпонентных ситуаций.

Чем компенсируется у человека очевидное преимущество компьютера в быстродействии? Ответ один: в основном, организацией памяти. Постараемся выделить структурные свойства памяти, которые, по нашему мнению, необходимы для формирования вразумительных объяснений человеческого феномена обработки информации.

1. Превалирующей тенденцией в развитии интеллектуальных систем является совершенствование человеко-машинного взаимодействия, вплоть до достижения партнерского уровня отношений. Поэтому необходимо использовать в компьютерах естественные, свойственные человеку принципы моделирования сред, ситуаций, задач. Типы моделей у партнеров (человека и компьютера) должны быть одинаковыми. В жизнедеятельности человека важное значение имеют логико-лингвистические информационные модели, т.е. такие, в которых основные элементы — не числа и вычислительные операции, а имена и логические связи. Логико-лингвистические модели адекватно описываются естественно-языковыми конструкциями, и в этом одно из важнейших их достоинств в организации человеко-машинного интерфей-

са. В будущих компьютерах должны быть созданы условия для человеко-машинного решения задач в партнерском режиме, обеспечивающем переключение от компьютера к человеку и наоборот в процессе решения одной задачи. Такой режим можно организовать только согласованием типов информационных моделей, используемых партнерами. Логико-лингвистические модели — наиболее приемлемый тип моделей для такого согласования.

2. Структура памяти формируется одновременно с восприятием и под влиянием воспринимаемой и уже хранящейся информации. Она отображает воспринятую информацию. Структуризация информации — неотъемлемая функция памяти. К основным процессам структуризации относятся формирование ассоциативных связей путем выделения пересечений признаков представлений объектов, иерархическое упорядочение, классификация, формирование обобщенных логических признаков моделей классов (понятий). В реальных условиях восприятия информации часто нет возможности получить полную информацию об объекте сразу, одновременно (например, из-за неудачного ракурса или освещенности при восприятии зрительной информации). Поэтому процессы формирования памяти должны допускать возможность «порционной», по частям, достройки моделей объектов и их классов.

3. В различных процессах обработки информации объекты представляются одним из двух способов: именем (свернутое, конвергированное представление) или набором значений признаков (развернутое представление). Структура памяти должна обеспечивать удобный переход от одного представления к другому. Механизмы, обеспечивающие такой переход в нейросистеме человека при опознании и воспоминании, рассматриваются в работах Г.С. Воронкова и З.Л. Рабиновича [1].

Сформулируем приведенные тезисы в виде требований к организации памяти в интеллектуальных системах.

- Знания разных типов должны объединяться в иерархическую сетевую структуру, построенную на единых для всех видов знаний принципах.
- Структура, представляющая знания, должна обладать развитыми ассоциативными свойствами и таким образом обеспечивать выполнение разнородных поисковых операций.
- Сеть должна отражать иерархичность реальных сред и в связи с этим быть удобной для представления родовидовых связей и структур составных объектов.
- Обязательными функциями памяти должны быть формирование ассоциативных связей путем выделения пересечений признаков представлений объектов, иерархического упорядочения, классификации, формирования понятий. Эти функции должны выполняться одновременно с восприятием информации.
- В сети должен быть обеспечен удобный двунаправленный переход между свернутым и развернутым представлениями объектов.

## РАСТУЩАЯ ПИРАМИДАЛЬНАЯ СЕТЬ

Приведенным требованиям удовлетворяют РПС. Применение РПС для решения аналитических задач в различных областях науки и техники подтвердило их репутацию эффективного средства структуризации больших объемов данных, обеспечивающего не только высокие результаты при решении аналитических задач, но и, в отличие от нейронных сетей, формирование легко интерпретируемых обобщенных описаний закономерностей, характеризующих классы объектов.

Теория и практические применения РПС описаны во многих публикациях [2–6]. В этой статье приводятся несколько измененные правила формирования РПС, обеспечивающие их построение при вводе признаков описаний объектов по частям, и сопоставляются результаты применения РПС при выполнении анализа данных индуктивными методами и методами, в основе которых лежит вывод по аналогии.

*Растущей пирамидальной сетью* называется ациклический ориентированный граф, в котором нет вершин, имеющих одну входящую дугу. Вершины, не имеющие входящих дуг, называются *рецепторами*, остальные — *концепторами*. Подграф пирамидальной сети, содержащий вершину  $a$  и все вершины, от которых имеются пути к  $a$ , называется *пирамидой* вершины  $a$ . Вершины, входящие в пирамиду вершины  $a$ , образуют ее *субмножество*. Множество вершин, к которым имеются пути от вершины  $a$ , называется ее *супермножеством*.

В субмноестве и супермноестве вершины выделяются 0-субмножество и 0-супермножество, состоящие из тех вершин, которые связаны с ней непосредственно. При построении сети входной информацией служат наборы значений признаков, описывающих некоторые объекты (материалы, состояния агрегата, ситуации, болезни и т.п.). Рецепторы соответствуют значениям признаков. В различных задачах это могут быть имена свойств, отношений, состояний, действий, объектов или классов объектов. Концепторы соответствуют описаниям объектов в целом и пересечениям описаний.

В начальном состоянии сеть состоит только из рецепторов. Концепторы формируются в результате работы алгоритма построения сети. Алгоритм [2–5] предназначен для работы в ситуациях, когда признаковое описание каждого объекта полностью известно и вводится целиком. При появлении новых признаков, характеризующих объект, необходимо формировать новое полное описание объекта и заменять представляющую его пирамиду другой, соответствующей новому описанию. Однако, как уже отмечалось, в реальных ситуациях функционирования интегрального интеллектуального агента далеко не всегда возможно одновременное восприятие всех характеристик объекта. В этих случаях информация об объектах поступает по частям. Появляется необходимость изменить алгоритм построения сети таким образом, чтобы обеспечить возможность включения в существующие пирамиды объектов новых признаков по мере их появления, без замены пирамид в целом. Приведем описание измененного алгоритма.

При вводе признакового описания объекта рецепторы, соответствующие значениям признаков, входящим в описание, переводятся в *состояние возбуждения*. Процесс возбуждения распространяется по сети. Концептор

переводится в состояние возбуждения, если возбуждены все вершины его 0-субмножества. Рецепторы и концепторы сохраняют состояние возбуждения в течение выполнения всех операций достройки сети.

Пусть при вводе описания объекта  $F_a$  – подмножество возбужденных вершин 0-субмножества вершины  $a$ ;  $G$  — множество возбужденных вершин сети, не имеющих других возбужденных вершин в своих супермножествах.

Ввод новых вершин производится по следующим правилам.

**Правило 1.**

Если вершина  $a$  не возбуждена, и множество  $F_a$  содержит более одного элемента, то дуги, соединяющие вершины из множества  $F_a$  с вершиной  $a$ , ликвидируются, и в сеть вводится новый концептор, который соединяется заходящими дугами с вершинами множества  $F_a$  и исходящей дугой с вершиной  $a$ . Новая вершина находится в состоянии возбуждения.

Выполнение правила 1 иллюстрируется рис. 1 (I, II).

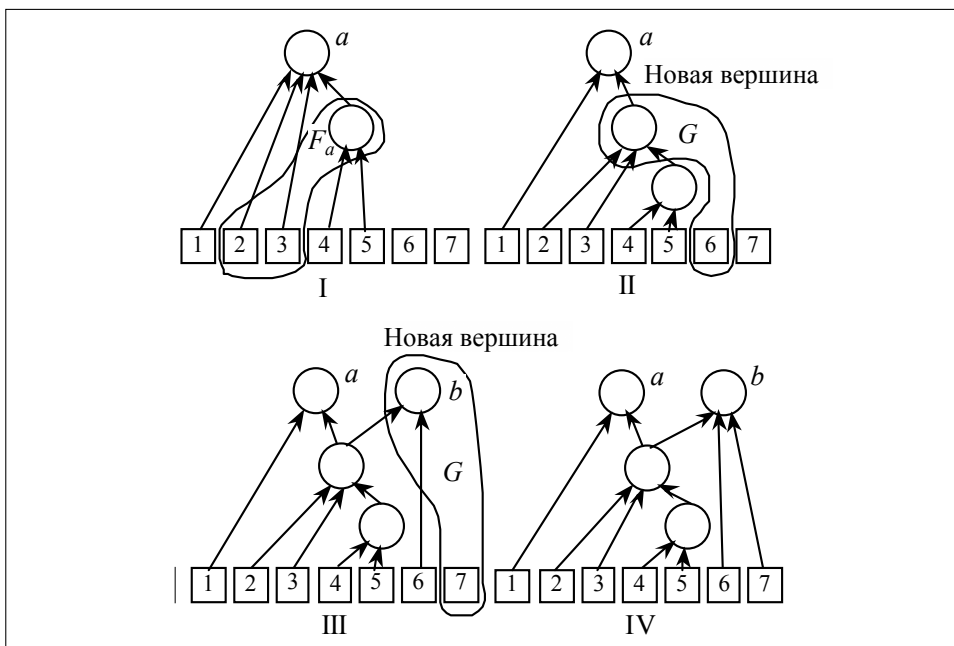


Рис. 1. Построение сети

Сеть II возникает после возбуждения в сети I рецепторов 2, 3, 4, 5.

Как следует из правила 1, условием ввода в сеть новой вершины является ситуация, когда некоторая вершина сети оказывается не полностью возбужденной (возбуждаются не все, но не менее двух вершин ее 0-субмножества). Новые вершины вводятся в субмножества не полностью возбужденных вершин.

После введения новых вершин во все участки, где удовлетворяется условие правила 1, выполняется правило 2 или 3.

### **Правило 2.**

Если множество  $G$  содержит более одного элемента и не содержит вершины, помеченной именем введенного объекта, к сети присоединяется новый концептор, который соединяется заходящими дугами со всеми вершинами множества  $G$ . Новая вершина находится в возбужденном состоянии. Выполнение правила 2 иллюстрируется рис.1(II,III). Сеть III возникает после возбуждения в сети II рецепторов 2, 3, 4, 5, 6.

### **Правило 3.**

Если множество  $G$  содержит более одного элемента, в том числе вершину, помеченную именем введенного объекта, то эта вершина соединяется входящими дугами с теми вершинами из множества  $G$ , которые не соединены с ней.

Выполнение правила 3 иллюстрируется рис.1(III,IV). Сеть IV возникает после возбуждения в сети III рецепторов 2, 3, 4, 5, 6, 7 при условии, что они соответствуют описанию объекта  $b$ .

В измененном алгоритме возможность включения в существующие пирамиды новых признаков обеспечивается правилом 3.

Пирамидальные сети удобны для выполнения различных операций ассоциативного поиска. Например, можно выбрать все объекты с заданным сочетанием значений признаков, прослеживая пути из вершины сети, соответствующей этому сочетанию. Для выборки всех объектов, описания которых пересекаются с описанием заданного объекта, достаточно проследить пути из вершин, образующих его пирамиду.

Алгоритм построения сети обеспечивает автоматическое установление ассоциативной близости между объектами по общим элементам их описаний. Все процессы построения сети при обработке одного описания локализуются в относительно небольшой ее части – пирамиде, соответствующей этому описанию.

Важным свойством пирамидальных сетей является их иерархичность, позволяющая естественным образом отображать структуру составных объектов и родовидовые связи.

Концепторы сети соответствуют сочетаниям значений признаков, определяющих конъюнктивные классы объектов. При включении возбужденных вершин в пирамиду объекта осуществляется привязка объекта к классам, определения которых представлены этими вершинами. Таким образом при построении сети формируются конъюнктивные классы объектов, т.е. осуществляется классификация без учителя. Классифицирующие свойства пирамидальной сети имеют большое значение для моделирования сред и ситуаций.

Переход от конвергированных представлений объектов (концепторов) к развернутым (наборам рецепторов) осуществляется путем просмотра пирамид в разных направлениях (сверху вниз и наоборот).

В РПС реализуются индуктивные процессы формирования обобщенных логических моделей классов объектов (понятий) [2–5]. Сформированное понятие представляется в сети ансамблем особо выделенных (*контрольных*) вершин. На основе анализа сети специальная процедура строит понятие в форме логического выражения.

В таблице приведен пример формирования понятий на основе анализа фрагмента обучающей выборки. Даны описания керамических материалов двух классов с использованием следующих признаков: М — материал, Т — тонкость порошка, С — его состав, РР — способ получения порошка, GP — условия получения образца при горячем прессовании, NoGP — другие условия получения образца, DU — особые условия изготовления образца, Por — пористость, Z — зернистость. Буквы и цифры обозначают значения соответствующих признаков.

Обучающая выборка

Объект	Класс	M	T	C	PP	GP	NoGP	DU	Por	Z
97	1	Al	2		SYN	2		2GP		
96	1	Al	2		SYN	2		1GP		
92	1	Al	2		SYN	2		2GP	1	
227	1	TiB	11	TiO-C	SYN		9		3	2
228	1	TiB	11	TiO-C	SYN		9		3	2
229	1	TiB	11	TiO-C	SYN		9		3	2
233	1	SiC	11	TiO-C	SYN		9		3	2
234	1	SiC	11	SiO-C	SYN		9		3	2
235	1	SiC	11	SiO-C	SYN		9		3	2
237	1	SiC	11	SiO-C	SYN		9		3	2
239	1	ZrB	11	ZrO-C	SYN		9		3	2
240	1	ZrB	11	ZrO-C	SYN		9		3	2
241	1	ZrB	11	ZrO-C	SYN		9		3	2
242	1	ZrB	11	ZrO-C	SYN		9		3	2
154	1	TiB	7	TiO-C	KRB	3			3	4
156	1	TiB	7	TiO-C	KRB	3			3	4
163	1	1AlO	1	AlO	SYN	1			4	
158	2	TiB	8	TiO-C	KRB	3			3	6
160	2	1AlO	1	AlO	SYN	1			1	
159	2	BC	1		SYN	1			1	
308	2	ZrB	11	ZrO-C	SYN		9			2

На рис. 2 показана соответствующая пирамидальная сеть после выполнения процесса формирования понятий. Контрольные вершины PP\_SYN, Por\_3, 239, 163 характеризуют класс 1, а 158, 308 и \$6 — класс 2.

Класс 1 описывается следующим логическим выражением, где  $\vee$ ,  $\wedge$ ,  $\neg$  — логические операции дизъюнкции, конъюнкции и отрицания:

$$PP\_SYN \wedge \neg\{T\_1 \wedge GP\_1\} \wedge \neg\{M\_ZrB \wedge C\_ZrO-C \wedge T\_11 \wedge NoGP\_9 \wedge Z\_2\} \vee$$

$$Por\_3 \wedge \neg\{T\_8 \wedge Z\_6 \wedge M\_TiB \wedge C\_TiO-C \wedge PP\_KRB \wedge GP\_3\} \vee$$

$$M\_ZrB \wedge C\_ZrO-C \wedge T\_11 \wedge PP\_SYN \wedge NoGP\_9 \wedge Por\_3 \wedge Z\_2 \vee$$

$$M\_1AlO \wedge T\_1 \wedge C\_AlO \wedge PP\_SYN \wedge GP\_1 \wedge Por\_4.$$

Логические выражения, определяющие различные классы объектов, объединяются в *кластерные базы данных* (КБД). КБД содержат информацию о группах объектов (кластерах), специфичных для исследуемой предметной области. На основе КБД решаются задачи классификации, диагностики и прогнозирования.

После того как понятие для некоторого класса объектов сформировано, проблемы прогнозирования и диагностики сводятся к задаче классификации. Классификация новых объектов выполняется путем сравнения их признаков с описанием понятия, определяющим класс прогнозируемых или диагностируемых объектов. Объекты можно классифицировать, вычисляя значение логических выражений, которые представляют соответствующие понятия.

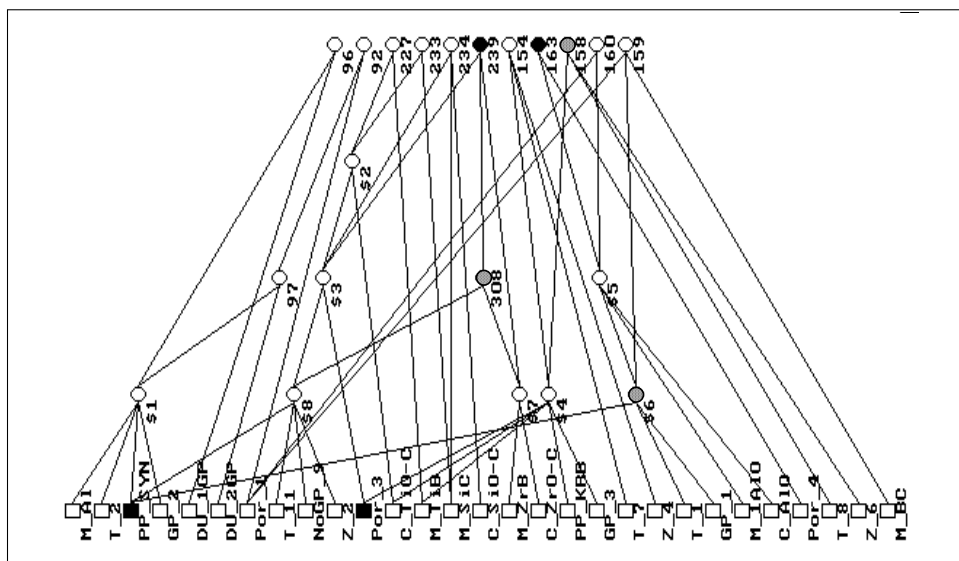


Рис. 2. Пример сети

Для лучшего понимания алгоритма формирования понятий дадим его геометрическую интерпретацию.

Каждой вершине сети, имеющей  $k$  рецепторов в своем субмножестве, соответствует  $(s - k)$ -мерная плоскость в  $s$ -мерном признаковом пространстве. Плоскость содержит все точки, представляющие объекты, при восприятии которых возбуждается эта вершина.  $(s - k)$ -мерные плоскости, соответствующие контрольным вершинам понятия  $Q_i$ , будем называть *зонами понятия*  $Q_i$ .

Для пирамидальных сетей справедливы следующие утверждения.

**Утверждение 1.** Зона некоторой вершины сети целиком входит в зоны вершин ее субмножества и содержит зоны вершин ее супермножества.

**Утверждение 2.** Точка, представляющая объект в признаковом пространстве, находится внутри области, образованной пересечением зон контрольных вершин, которые возбуждаются при восприятии этого объекта.

В результате работы алгоритма для каждого из сформированных понятий строится область из зон признакового пространства, содержащая все

точки объектов соответствующего класса и не содержащая ни одной точки, представляющей объекты других классов. Она аппроксимирует распределение объектов соответствующего класса и состоит из линейных элементарных областей (гиперплоскостей), а ограничивающая ее поверхность является кусочно-линейной. Следовательно, алгоритм осуществляет кусочно-линейное разделение объектов, соответствующих разным понятиям.

Описанный метод обеспечивает решение аналитических проблем классификации, диагностики и прогнозирования на основе логических моделей классов объектов. Модель отображает зависимости исследуемого класса от сочетаний значений признаков, т.е. дает возможность учесть эффект совместного влияния нескольких признаков.

Существенной особенностью метода формирования понятий в пирамидальных сетях является возможность введения в понятия так называемых исключаяющих признаков, которые не принадлежат объектам исследуемого класса. Таким образом сформированные понятия имеют более компактную логическую структуру, что в принципе позволяет увеличить точность диагноза или прогноза. В логическом выражении исключаяющие признаки представлены переменными с отрицанием.

Все поисковые операции пирамидальной сети ограничиваются сравнительно малым фрагментом сети, который содержит пирамиду объекта и вершины, непосредственно связанные с ней. Это создает принципиальную возможность решать практические аналитические проблемы на основе больших объемов данных.

В пирамидальной сети информация хранится в виде отображения в структуре сети. Информация об объектах и классах объектов представлена ансамблями вершин (пирамидами), распределенными по всей сети. Внесение новой информации вызывает перераспределение связей между вершинами сети, т.е. изменение ее структуры.

Конечно, в полной мере достоинства пирамидальных сетей проявляются при их физической реализации, допускающей параллельное распространение сигналов по сети. Важным свойством сети как средства хранения информации является то, что параллельное распространение сигналов сочетается в ней с возможностью параллельного приема сигналов на рецепторы.

Существует аналогия между основными процессами в РПС и нейронных сетях. Главное преимущество РПС: ее структура формируется полностью автоматически в зависимости от вводимых данных. В результате достигается оптимизация представления информации путем адаптации структуры сети к структурным особенностям данных. Причем, в отличие от нейронных сетей, эффект адаптации достигается без введения априорной избыточности сети. Процесс обучения не зависит от ее predetermined конфигурации. Недостатком нейронных сетей по сравнению с РПС является то, что выделенные в них обобщенные знания не могут быть явно представлены в виде правил или логических выражений. Это затрудняет их интерпретацию и понимание человеком.



## ЭКСПЕРИМЕНТЫ И ПРИМЕНЕНИЕ

Программный комплекс, используемый для проведения экспериментов и решения прикладных задач, состоит из трех систем.

1. CONFOR — реализует индуктивные методы.
2. ANALOGY — использует вывод по аналогии.
3. DISCRET — с ее помощью признаки, заданные в числовых шкалах, преобразуются в номинальные. Эта задача называется задачей дискретизации. Дискретизация выполняется на шкалах числовых признаков сравнением распределений объектов обучающей выборки, принадлежащих различным классам.

Вывод по аналогии является основой методов, суть которых состоит в анализе объектов обучающей выборки, наиболее «похожих» на исследуемый объект. Исследуемый объект считается принадлежащим классу, чьи объекты соответствуют в пространстве признаков точкам, расположенным наиболее близко к точке, представляющей исследуемый объект (метод  $K$  ближайших соседей). Проблема поиска аналогии должна решаться вместе с вопросами организации памяти, обеспечивающей установление аналогии объектов.

В свете современных воззрений это должна быть память коннекционистского типа, допускающая параллельное выполнение поисковых операций и отражающая в своей структуре семантические пересечения блоков информации. Этим требованиям отвечают растущие пирамидальные сети. Алгоритм построения сети, работающий при вводе описаний новых объектов, одновременно является алгоритмом поиска в сети аналогов нового объекта, имеющих общие с ним фрагменты описаний. Степень подобия объектов оценивается мощностью пересечения их атрибутивных описаний.

Комплекс прошел длительное испытание временем. Типичные прикладные задачи, для решения которых использовался комплекс: прогнозирование новых химических соединений и материалов с заданными свойствами, солнечной активности, нарушений в работе сложных агрегатов, а также прогнозирование в генетике, геологии, медицинской и технической диагностика.

Сравнение методов прогнозирования на основе индуктивного вывода и вывода по аналогии проводилось на задачах прогнозирования существования неорганических соединений с заданными свойствами.

В качестве обучающей выборки использовались таблицы, содержащие атрибутивные описания двойных, тройных и четверных систем химических элементов, образующих и не образующих химические соединения. Обучающие выборки для двойных, тройных и четверных систем содержали соответственно 1333, 4278 и 4963 описания, а экзаменационные выборки — 692, 2156 и 2536 описаний. Каждый химический элемент описывался 87 признаками. Описания двойных, тройных и четверных систем состояли соответственно из 174, 261 и 348 признаков. Была достигнута достаточно высокая точность прогноза — от 97,88 до 99,86%.

На основании результатов экспериментов можно сделать следующие выводы:

1. Методы, основанные на аналогии с предварительным переводом числовых признаков в номинальные (система DISCRET), более просты в использовании и дают хорошие результаты прогноза при условии применения представительных обучающих выборок в случаях, когда области распределения объектов разных классов в пространстве признаков компактны. Без предварительного перевода признаков в номинальные реализация методов существенно усложняется из-за необходимости согласования единиц измерения для признаков, заданных в различных шкалах.

2. В методах, основанных на аналогии, результаты классификации зависят от параметра, определяющего размеры анализируемого окружения точки, представляющей объект в признаковом пространстве (например, число  $K$  в методе  $K$  ближайших соседей). Естественно выбирать  $K$  таким образом, чтобы обеспечить наилучшую классификацию объектов экзаменационной выборки. В этом случае параметр  $K$  может идеально разделять экзаменационную выборку и, тем не менее, давать плохие результаты при классификации других объектов.

3. Использование методов, содержащих индуктивный вывод, снижает требования к обучающей выборке. Однако в отличие от методов, основанных на аналогии, индуктивные методы гораздо чаще приводят к появлению неопределенного прогноза (ответа типа «не знаю»). Это происходит, в частности, в случаях, когда исследуемый объект одновременно соответствует закономерностям, характеризующим различные классы. Работу системы CONFOR, реализующей методы, которые содержат индуктивный вывод, можно сравнить с поведением серьезного исследователя, «не бросающего слов на ветер». Работая в очень сложном «зашумленном» признаковом пространстве, он часто отвечает «не знаю», однако дает прогноз с большой степенью точности.

4. Иногда важным достоинством индуктивных методов прогнозирования является создание обобщенной модели исследуемого класса объектов, которая представляется в виде логического выражения, удобного для интерпретации человеком.

## **ЗАКЛЮЧЕНИЕ**

Растущая пирамидальная сеть является сетевой памятью, автоматически настраиваемой на структуру входной информации. В отличие от сетей нейроподобных элементов эффект адаптации достигается без введения априорной избыточности сети.

Исследования, выполненные на сложных данных большого объема, показали высокую эффективность применения растущих пирамидальных сетей для решения аналитических задач. Такие качества, как простота внесения изменений, совмещение процессов ввода информации с ее классификацией, обобщением и выделением существенных признаков, высокая ассоциативность, делают растущие пирамидальные сети важным компонентом прогнозирующих и диагностических систем.

## ЛИТЕРАТУРА

1. *Воронков Г.С. Рабинович З.Л.* Естественная среда памяти и мышления: модельное представление //Тр. междунар. конф. «Знание — Диалог — Решение — 2001». — СПб: Изд. Сев.-Зап. политехн. ин-та, 2001. — С. 110–115.
2. *Гладун В.П.* Партнерство с компьютером. Человеко-машинные целеустремленные системы. — Киев: Port-Royal, 2000. — 118 с.
3. *Гладун В.П. Ващенко Н.Д.* Локально-статистические методы формирования знаний // Кибернетика и системный анализ. — 1995. — № 2. — С. 62–74.
4. *Гладун В.П.* Планирование решений. — Киев: Наук. думка, 1987. — 167 с.
5. *Gladun V.P. and Vashchenko N.D.* Analitical processes in pyramidal networks // Intern. Journal on Information Theories and Applications. FOI-COMMERCE, Sofia. — 2000. — 7, № 3. — С. 103–109.
6. *Kiselyova N., Gladun V., Vashchenko N.* Computational Materials Design Using Artificial Intelligence Methods // Journal of Alloys and Compounds. — 1998. — № 279. — P. 8–13.

*Поступила 18.06.2003*