ОБНАРУЖЕНИЕ ЭМПИРИЧЕСКИХ ЗАКОНОМЕРНОСТЕЙ С ПОМОЩЬЮ ЭВМ
(Вычислительные системы)
1984 год Выпуск 102

УДК 007:62-50

МЕТОДЫ ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ НА ОСНОВЕ РАСТУЩИХ ПИРАМИДАЛЬНЫХ СЕТЕЙ

Л.В.Хоменко

Важным условием дальнейшего повышения эффективности научных исследований является автоматизация обнаружения закономерностей при анализе экспериментальных данных. В статье рассматриваются методы обнаружения закономерностей на основе индуктивного обучения с использованием структур данных в виде растущих пирамидальных сетей [I,гл.2].

З а д а ч а ф о р м и р о в а н и я п о н я т и й. Поиск закономерностей осуществляется для класса исследуемых объектов V и $\bar{V}$ $(V \cap \bar{V} = \emptyset)$ в рамках задачи построения обобщенного определения классов объектов (задачи формирования понятий). Общая постановка задачи формирования понятий для объектов разных классов, заданных конечными наборами значений признаков, приводится в работе [I, гл.I]. Пусть $L=\{a\}$ - множество объектов случайной и независимой обучающей выборки, $L^V=L\cap V$, $L^{\bar{V}} = L\cap \bar{V}$. Объекты множеств V, $\bar{V}$ могут описываться признаками, измеренными в разнотипных шкалах. Описание объектов представляется конъюнкцией истинных на объекте (равных I) одноместных предикатов.

Понятия о множествах объектов V и $\bar{V}$ выражаются в виде логических функций Π^V и $\Pi^{\bar{V}}$:

$$\Pi^V = \prod_{i=1}^{p_V} S_i^V = \begin{cases} 1 & \text{при } a \in L^V, \\ 0 & \text{при } a \in L^{\bar{V}}, \end{cases}$$

$$\Pi^{\bar{V}} = \prod_{i=1}^{p_{\bar{V}}} S_i^{\bar{V}} \quad \begin{cases} 0 & \text{при } a \in L^V, \\ 1 & \text{при } a \in L^{\bar{V}}, \end{cases}$$

где S_i^V, $S_i^{\bar{V}}$ – конъюнкции предикатов и отрицания конъюнкций предикатов. Надо для L^V и $L^{\bar{V}}$ в совокупности получить минимальное или близкое к нему число ($p+t \ll |L|$) логических функций S^V и $S^{\bar{V}}$, или з а к о н о м е р н о с т е й. Относительное число объектов из $L^V(L^{\bar{V}})$, на которых $S^V(S^{\bar{V}})$ равна I, называется э ф ф е к т и в н о с т ь ю закономерности $S^V(S^{\bar{V}})$.

С т р у к т у р ы д а н н ы х д л я р е ш е н и я з а д а ч и ф о р м и р о в а н и я п о н я т и й. Решение указанной задачи может быть реализовано с использованием различных структур данных. Проведенный анализ структур данных для задач формирования понятий показал целесообразность выбора структур данных в виде растущих пирамидальных сетей, благодаря их свойствам ассоциативности и иерархичности.

Растущая пирамидальная сеть – ациклический ориентированный граф, в котором нет вершин со степенью захода I. Она строится на основании алгоритма, описанного в работе [I,гл.2]. Объекты обучающей выборки представлены в ней своими пирамидами. П и р а м и д а о б ъ е к т а – сураф растущей пирамидальной сети (ориентированный подграф), включающий вершину с 0-степенью исхода (г л а в н у ю вершину пирамиды объекта) и все вершины, из которых достижима главная вершина. Каждой вершине пирамидальной сети соответствует предикат или конъюнкция предикатов, равных I на каком-либо объекте или группе объектов обучающей выборки. Вершинам с 0-степенью захода (р е ц е п т о р а м) пирамиды объекта соответствуют предикаты из описания данного объекта. Остальные вершины пирамиды объекта называются ассоциативными вершинами. С у б м н о ж е с т в о вершины $\alpha(SB_\alpha)$ – множество, включающее все вершины, из которых достижима вершина α, в том числе α. 0-субмножество вершины α (SB_α^0) – вершины SB_α, связанные непосредственно с α. Слой 0-го порядка состоит из главных вершин. С л о й m-го порядка составляют вершины 0-субмножества вершин (m-1)-го порядка ($1 \le m \le n$, n – максимальная длина пути от рецепторов к главным вершинам пирамид объектов множества L). С у п е р м н о ж е с т в о вершины $\alpha(SP_\alpha)$ – множество всех вершин, которые достижимы из вершины α. 0-супермножество вершины $\alpha(SP_\alpha^0)$ – это вершины SP_α, связанные непосредственно с α. Э л е м е н т а р н ы й с у г р а ф с е т и – сураф, включающий не главную вершину α со степенью исхода больше I и SP_α^0.

Задача формирования понятий на основе растущей пирамидальной сети, построенной для объектов обучающей выборки, состоит в выделении на ней минимального или близкого к нему числа помеченных вершин графа (К-вершин класса V и $\bar{V}$). На основе данного множества К-вершин и некоторого правила распознавания должно осуществляться правильное распознавание объектов множества L. Множество выделенных с помощью алгоритма формирования понятий К-вершин соответствует множеству обнаруженных закономерностей.

Введем важную структурную характеристику вершин пирамидальной сети объектов множества L - степень противоречивости. Алгоритмы формирования понятий на основе сети обеспечивают поиск эффективных закономерностей в случае учета степени противоречивости вершин, а также весовых характеристик вершин. Весовые характеристики вершин будут даны при описании алгоритма формирования понятий. Вершина сети п р о т и в о р е ч и в а, если она одновременно принадлежит пирамидам объектов из множеств L^V и $L^{\bar{V}}$. Противоречивой вершине в сети соответствует предикат или конъюнкция предикатов, которые входят в описание объектов класса V и $\bar{V}$. Наличие большого числа противоречивых вершин в сети свидетельствует о сложном расположении точек, отображающих объекты разных классов в пространстве признаков. С т е п е н ь п р о т и в о р е ч и в о с т и вершины α ($ST(\alpha)$) - число противоречивых вершин из SP_α. Степень противоречивости вершин сети для фиксированного $|L|$ может изменяться в пределах от 0 до $|L|-1$. В сети могут быть противоречивыми все вершины, кроме главных. Под степенью противоречивости сети будем понимать среднюю степень противоречивости вершин сети, не являющихся главными. Степень противоречивости является интегральной характеристикой сложности сети обучающей выборки.

В разработанном алгоритме формирования понятий А2 [2,3] реализуется процесс сведения анализа сложной (с большой степенью противоречивости) сети объектов обучающей выборки к анализу простых (с малой степенью противоречивости) суграфов сети, в частности, к анализу элементарных суграфов сети.

А л г о р и т м ы ф о р м и р о в а н и я п о н я т и й н а о с н о в е р а с т у щ и х п и р а м и д а л ь н ы х с е т е й. Приведем краткое описание алгоритма А2. Алгоритм А2 реализует итеративный процесс обобщения. На каждом шаге алгоритма строятся приближения к понятиям $П^V$ и $П^{\bar{V}}$. Начальными приближениями являются логические функции, включающие описания всех объектов L^V и $L^{\bar{V}}$ соответственно. Процесс обобщения на сети реа-

лизуется путем выделения на каждом шаге множеств К-вершин, обеспечивающих правильное распознавание всех объектов множества L. Число К-вершин от шага к шагу уменьшается. В качестве начального множества К-вершин выбираются главные вершины пирамид объектов обучающей выборки. На каждом шаге происходит замена групп К-вершин слоя (m-1)-го порядка отдельными К-вершинами слоя m-го порядка ($1 \le m \le n$), т.е. осуществляется объединение и спуск К-вершин от главных вершин к рецепторам по слоям. Порядок выбора вершин для анализа и замены в пределах слоя определяется оценкой значимости этих вершин для распознавания объектов множества L - функцией δ: $\delta(\alpha) = |b_0^V(\alpha) - b_0^{\bar{V}}(\alpha)|$, где $b_0^V(\alpha)$ ($b_0^{\bar{V}}(\alpha)$) - число К-вершин класса V ($\bar{V}$) в SP_α^0. Вершины рассматриваются в порядке от δ_{max} к δ_{min}. Вершина α становится К-вершиной того класса объектов, для которого она более характерна. При этом действуют правила замены:

1) вершина α выделяется в качестве К-вершины класса V при $b_0^V(\alpha) \ge b^{\bar{V}}(\alpha)$;

2) вершина α выделяется в качестве К-вершины класса $\bar{V}$ при $b_0^{\bar{V}}(\alpha) > b^V(\alpha)$;

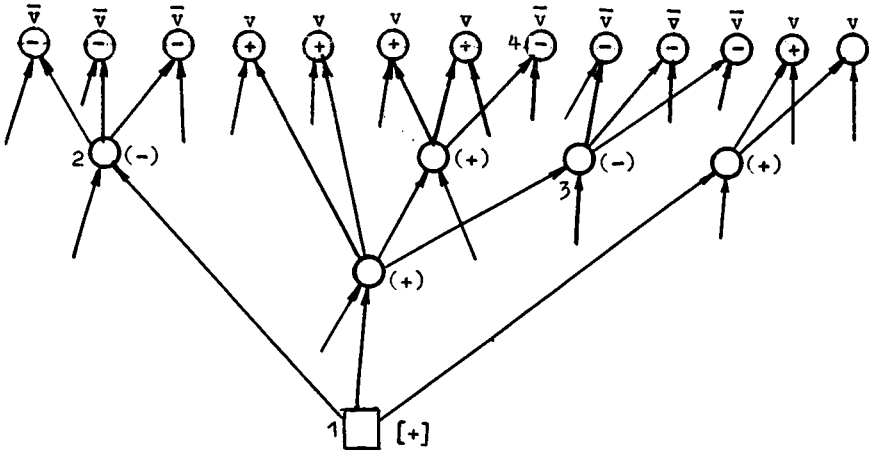
3) в остальных случаях вершина α не выделяется в качестве К-вершины, где $b^V(\alpha)$($b^{\bar{V}}(\alpha)$) - число К-вершин класса V($\bar{V}$) в SP_α. Числа $b_0^V, b_0^{\bar{V}}, b^V, b^{\bar{V}}$ являются весовыми характеристиками вершин пирамидальной сети. Таким образом, благодаря объединению и замене общее число К-вершин уменьшается, приближаясь к минимальному числу К-вершин. В алгоритме А2 анализ по слоям вместе с процессом объединения и замены К-вершин соответствует переходу от сравнения описаний отдельных объектов к сравнению описаний все более крупных групп объектов. Диапазон анализа на каждом шаге становится все более "узким", что позволяет избежать большого числа просмотров пирамид объектов и эффективно решать проблему переборов.

Число выделенных алгоритмом А2 К-вершин равно числу закономерностей. Логические выражения закономерностей можно получить с помощью алгоритма построения логического выражения [I, гл.3]. К-вершина α, в SP_α которой не содержится К-вершин противоположного класса, выразится в форме конъюнкции предикатов. Подмножество К-вершин, включающее К-вершину α и К-вершины противоположного класса из SP_α, выразится в форме конъюнкции предикатов и отрицаний конъюнкций предикатов.

Алгоритм А2 использует степень противоречивости неявно (без прямого подсчета). Разработана модификация алгоритма А2 - алгоритм А3 с подсчетом степени противоречивости вершин. По алгоритму А3 вершины каждого слоя упорядочиваются и анализируются в соответствии с минимальным значением функции $\delta T(\alpha)$ и максимальным значением функции $\delta(\alpha)$.

А н а л и з. Р е а л и з а ц и я и п р и м е н е н и е. Алгоритмы формирования понятий АI [I,гл.3] и А2 и А3 отличаются принципами анализа объектов множества L. В алгоритме АI реализуется принцип последовательного, а в алгоритмах А2, А3 - принцип параллельного анализа. Последовательность анализа в АI позволяет использовать его в режиме "реального времени", когда необходимо формировать понятие путем последовательной корректировки при вводе нового объекта обучения. Параллельность анализа с использованием всех объектов множества L в алгоритмах А2, А3 позволяет обнаруживать эффективные закономерности [2,3].

Проведен анализ алгоритмов АI, А2 и А3 на разновидностях сетей (т.е. сети с различными структурными характеристиками вершин). Назовем ассоциативную вершину или рецептор со степенью исхода больше I вершиной фокусом. О с н о в н о й ф о к у с - это фокус, субмножество которого не содержит вершин с такими же свойствами. Назовем растущую пирамидальную сеть объектов множества L о д н о ф о к у с н о й, если в ней содержится один основной фокус, и а л ь т е р н а т и в н о й, если число основных фокусов больше I. Произвольная сеть является однофокусной или альтернативной. Процесс обобщения на большинстве альтернативных сетей можно представить как композицию процессов обобщения на однофокусных сетях. А процесс обобщения на однофокусной сети сводится к процессу обобщения на дереве, растущем из основного фокуса данной сети (базовом дереве). На рисунке изображено базовое дерево однофокусной сети, растущее из основного фокуса - вершины I. На этом примере демонстрируется работа алгоритма А2. Через v $(\bar{v})$ обозначены главные вершины пирамиды объекта класса $V(\bar{V})$. Рецепторы, которые не выделены в качестве К-вершин, на рисунке не изображены, показаны только дуги, исходящие из данных вершин. К-вершины класса $V(\bar{V})$, выделенные на различных шагах работы алгоритма А2, обозначены знаками +(-) в вершинах, а также в круглых и квадратных скобках рядом с вершинами. Окончательное множество К-вершин содержит минимальное число К-вершин и включает К-вершину I класса V, К-вершины 2,3,4 класса $\bar{V}$.



Для оценки эффективности алгоритма на основе растущей пирамидальной сети введем функцию φ, равную числу выделенных с помощью алгоритма формирования понятий К-вершин. Очевидно, что чем меньше φ, тем больше эффективность обнаруженных закономерностей. Будем считать, что алгоритм формирования понятий А принимает оптимальное решение в произвольной вершине α сети, если: 1) при выборе α в качестве К-вершины принимаются во внимание только К-вершины SP^{0}_{α} элементарного суграфа с фокусом в α; 2) общее число выделенных К-вершин в элементарном суграфе минимально, исходя из правила замены К-вершин алгоритма А2. Будем считать, что алгоритм формирования понятий А принимает о п т и м а л ь н о е р е ш е н и е на произвольной сети, если число выделенных с помощью алгоритма К-вершин минимально ($\varphi_A = \varphi_{min}$), и является достаточным для правильного распознавания объектов множества L. Для алгоритмов А2, А3 принятие оптимального решения на сети эквивалентно принятию оптимального решения на всех вершинах базовых деревьев, соответствующих данной сети. Будем считать, что алгоритм формирования понятий А принимает частично-оптимальное (не оптимальное) решение на произвольной сети, если А принимает оптимальное решение на большинстве (меньшинстве) вершинах базовых деревьев.

Справедливы следующие утверждения.

УТВЕРЖДЕНИЕ I. Число однофокусных сетей для фиксированного $|L|$, на которых алгоритм А2 принимает оптимальное или частично-оптимальное решение, больше числа однофокусных сетей для фиксированного $|L|$, на которых алгоритм А2 принимает не оптимальное решение.

УТВЕРЖДЕНИЕ 2. На большинстве разновидностей однофокусных сетей для фиксированного $|L|$, $\varphi_{A2} \leq \varphi_{A1}$, $\varphi_{A3} \leq \varphi_{A1}$.

Показано, что на большинстве разновидностей альтернативных сетей для фиксированного $|L|$, $\varphi_{A2} \leq \varphi_{A1}$, $\varphi_{A3} \leq \varphi_{A1}$.

УТВЕРЖДЕНИЕ 3. На всех однофокусных сетях алгоритм А3 принимает оптимальное решение, т.е. $\varphi_{A3} = \varphi_{min}$.

ДОКАЗАТЕЛЬСТВО данных утверждений основано на анализе свойств алгоритмов на разновидностях сетей, а также на сравнении (по числу решений) систем линейных неравенств, которым удовлетворяют весовые характеристики вершин разновидностей сетей.

Таким образом, алгоритмы А2, А3 обнаруживают на большинстве разновидностей сетей более эффективные закономерности по сравнению с алгоритмом А1.

Преимущества алгоритмов А2, А3 по сравнению с другими алгоритмами формирования понятий путем обобщения по признакам (алгоритмами Ханта, Бонгарда, Лбова [4-6]) состоят в возможности их реализации в семантической сети в виде сети, в сокращении переборов больших объемов данных, в обнаружении в ряде случаев более эффективных закономерностей.

Алгоритм А2 реализован в программной системе индуктивного формирования понятий "Анализатор-ЕС", разработанной в Институте кибернетики им. В.М.Глушкова АН УССР. Система применялась для обнаружения закономерностей и прогнозирования существования химических соединений с определенными свойствами, для выбора оптимальных параметров электрогидравлических установок, обеспечивающих заданный процесс формирования пробоя, для описания и прогнозирования существования классов рецептов резиносмесей с заданными свойствами.

Литература

I. ГЛАДУН Б.П. Эвристических поиск в сложных средах. - Киев: Наукова думка, 1977. - 166 с.

2. ХОМЕНКО Л.Б. Об одном алгоритме индуктивного формирования понятий на основе растущих пирамидальных сетей. -В кн.: Теоретические вопросы проектирования вычислительных систем. Киев, 1980, с. 33-47.

3. БОНДАРОВСКАЯ В.М., ХОМЕНКО Л.В. Опыт использования результатов психологического исследования для совершенствования системы искусственного интеллекта. -Кибернетика, 1979, № 3, с. 77-81.

4. ХАНТ Э., МАРИН Дж., СТОУН Ф. Моделирование процесса формирования понятий на вычислительной машине. -М.: Мир, 1970.-301 с.

5. БОНГАРД М.М. Проблема узнавания. -М.: Наука,1967. -320 с.

6. ЛБОВ Г.С. Методы обработки разнотипных экспериментальных данных. -Новосибирск: Наука, 1981. - 160 с.

Поступила в ред.-изд.отд.
22 марта 1984 года