А Н А Л И З   Р А З Н О Т И П Н Ы Х   Д А Н Н Ы Х
(Вычислительные системы)
1983 год Выпуск 99

УДК 519.95

КОНЪЮНКЦИЯ - ПРОГРАММА ПОСТРОЕНИЯ ЛОГИЧЕСКИХ ОТДЕЛИТЕЛЕЙ

В.Г. Устюжанинов

Пусть $x_1,\dots,x_l$ - система порядковых признаков; $Z_1,\dots,Z_l$ - конечные множества их градаций; M_1, M_2 - конечные непересекающиеся множества объектов I и 2-го классов, описываемых этими признаками. Образуем список $\mathcal{P}$ всевозможных предикатов P вида $x \le a$ и $x > a$, где $a \in Z_1 \cup \dots \cup Z_l$. Пусть k - конъюнкция некоторых предикатов из $\mathcal{P}$ и $M_1(k)$, $M_2(k)$ - множества объектов I и 2-го классов, на которых она истинна. Требуется построить такую дизъюнкцию $D = k_{\nu_1} \vee \dots \vee k_{\nu_s}$, состоящую из конъюнкций $k_{\nu_1},\dots,k_{\nu_s}$ предикатов из $\mathcal{P}$, что

$$M_1(D) = M_1, \tag{1}$$

число s минимально и для всех $i = 1,\dots,s$ имеет место

$$M_2(k_{\nu_i}) = \emptyset \tag{2}$$

Эта задача универсально переборна, и точное ее решение [I] требует непомерно большого машинного времени. Ниже излагается градиентная схема, дающая приближенное решение задачи.

Пусть $Q(\alpha) = \{k/\alpha \in M_1(k),\ M_2(k) = \emptyset\}$. Алгоритм работает в два этапа: на 1-м - выделяется такой список конъюнкций $K \subseteq \bigcup_{\alpha \in M_1} Q(\alpha)$, что $K \cap Q(\alpha) \neq \emptyset$ для всех $\alpha \in M_1$. Назовем этот список результирующим; на втором, - используя список K и градиентную процедуру С.В.Яблонского, И.А.Чегис, строится дизъюнкция D, удовлетворяющая (I).

Метод получения списка K заключается в преобразовании списка конъюнкций, именуемого буфером. Зафиксируем объект $\alpha \in M_1$. Пусть уже построен буфер, состоящий из некоторых конъюнкций, истинных на

α, и k_i - одна из конъюнкций буфера. Составим полный список предикатов $\mathcal{P}(\alpha)$, истинных на α и ложных по крайней мере на одном объекте из $M_2 \setminus M_2(k_i)$. Для каждого $P \in \mathcal{P}(\alpha)$ вычислим величину $B(P) = \text{ВЕС} \cdot \|M_1(k_i \& P)\| + (1 - \text{ВЕС}) \cdot \|M_2 \setminus M_2(k_i \& P)\|$, где $0 \le \text{ВЕС} \le 1$. Среди $P \in \mathcal{P}(\alpha)$ найдем P^*, для которого $B(P^*) = \max B(P)$.

Пусть $0 \le \text{ПР} \le 1$ и $P_{\mu_1}, \ldots, P_{\mu_l}$ - те предикаты, для которых справедливо $B(P_{\mu_j}) \ge B(P^*) \cdot \text{ПР}$. Исключим из буфера конъюнкцию k_i и пополним его конъюнкциями $k_i \& P_{\mu_1}, \ldots, k_i \& P_{\mu_l}$. Из полученного буфера извлечем конъюнкции, удовлетворяющие (2), и поместим их в результирующий список K. Если после этого буфер оказался пуст, то переходим к следующему объекту из M_1. В противном случае процесс повторяется.

Программа КОНЪЮНКЦИЯ, реализующая этот метод, имеет следующие механизмы, предназначенные для экономии памяти и машинного времени. Структура буфера древообразна. Из конъюнкций, подлежащих введению в буфер, включаются в буфер только те, для которых $M_1(k_i \& P_{\mu_t}) \not\subseteq M_1(k_j)$, где k_j - произвольная конъюнкция из образованного к этому моменту результирующего списка. Назовем конъюнкцию $k_j \in K$ тупиковой, если удаление из нее любого предиката приводит к конъюнкции, нарушающей требование (2). В каждой $k_j \in K$ выделяется тупиковая конъюнкция, которой она заменяется в списке K.

Программа КОНЪЮНКЦИЯ успешно применялась для решения задач распознавания образов в органической химии, горном деле, геологии. В частности, в геологии требовалось получить простейшее правило, отделяющее класс месторождений с промышленным содержанием полезного ископаемого от класса месторождений, не представляющих интереса для промышленности. Методами многомерного статистического анализа эта задача решалась И.Л.Добрецовым и др. в [2]. Ими отмечается, что полученные правила для распознавания месторождений представляют очень трудную задачу, которая обычными методами решается плохо. В этой связи интересно было посмотреть как справится с этой задачей программа КОНЪЮНКЦИЯ. Подробное изложение результатов этого эксперимента дается в [3]. Здесь же будем предельно кратки. Информация по задаче, любезно предоставленная лабораторией геологии и геофизики Новосибирского государственного университета, состояла из 46 объектов обучающей выборки. Объекты были разбиты на два класса: 14 - в первом классе и

32 - во втором. Каждый объект описывался 14 признаками. Программой было построено правило, состоящее всего из двух конъюнкций и безошибочно разделяющее классы объектов обучающей выборки. Для оценки эффективности этого правила было образовано множество из 34 контрольных объектов. Все они относились ко второму классу.

В отличие от объектов обучающей выборки контрольные объекты для построения отделяющего правила не использовались. После применения к контрольным объектам отделяющего правила, 4 из них были отнесены к первому классу, а остальные 30 - ко второму. По мнению геологов, полученный процент ошибки в 0.12 для решавшейся задачи - рекордно низкий.

Существуют другие программы построения логических отделителей, например, описанные в [4]. Сравнение эффективности этих программ с программой, представленной в настоящей заметке, - дело будущего.

Литература

1. ЖУРАВЛЁВ Ю.И. Об отделимости вершин n-мерного единичного куба. - Труды МИ АН СССР, 1958, т. 51, с. 143-157.

2. ДОБРЕЦОВ Н.Л., ХАРЬКИВ А.Д., ШЕМЯКИН М.Л. Применение многомерного статистического анализа для решения прогнозных задач на примере алмазоносности кимберлитов. - Геология и геофизика. 1966, № 8, с. 15-22.

3. УСТЮЖАНИНОВ В.Г. Исследование алгоритмов, решающих дискретные задачи распознавания образов: Дис. на соиск. учен. степ. канд. физ.-мат. наук. - Москва,1980.

4. ЛБОВ Г.С., КОТЮКОВ В.И., МАШАРОВ Ю.П. Метод обнаружения логических закономерностей на эмпирических таблицах. -В кн.: Эмпирическое предсказание и распознавание образов (Вычислительные системы, вып. 67). Новосибирск, 1976, с. 29-42.

Поступила в ред.-изд.отд.
10 марта 1982 года