Классификация по прецедентам и поиск в данных частых элементов

Н.А. Драгунов
Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Москва, Россия
nikitadragunovjob@gmail.com

Е.В. Дюкова
Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Москва, Россия
edjukova@mail.ru

А.П. Дюкова
Санкт-Петербургский научно-исследовательский институт эпидемиологии и микробиологии им. Пастера
Санкт-Петербург, Россия
anastasia.d.95@gmail.com

***Аннотация*—Исследуются корректные логические классификаторы, при конструировании которых используются как основные идеи хорошо известного алгоритма «Кора», так и алгоритмов вычисления оценок. Предлагается модификация этапа обучения, позволяющая в определённых случаях сократить временные затраты без потери качества классификации. Приводятся результаты экспериментов на модельных и реальных данных и новые теоретические оценки сложности обучения классификаторов типа «Кора».**

Ключевые слова*— *классификация по прецедентам, корректный логический классификатор типа «Кора», представительный элементарный классификатор, частый элементарный классификатор, тупиковое покрытие целочисленной матрицы.

1. Введение

В работе исследуются вопросы сокращения временных затрат на этапе обучения корректных классификаторов, базирующихся на применении методов дискретной математики. Одна из первых моделей классификаторов типа «Кора» предложена в [1]. В дальнейшем подход развивался в ряде работ ([2]– [4] и др.).

Обучение рассматриваемых классификаторов традиционно заключается в построении специальных фрагментов признаковых описаний прецедентов. Каждый такой фрагмент, называемый тупиковым представительным элементарным классификатором, позволяет отличать порождающий его объект от прецедентов из других классов, и при этом является в некотором смысле минимальным. Требование минимальности приводит к необходимости решать задачу построения тупиковых покрытий целочисленной матрицы, которая относится к числу труднорешаемых дискретных задач. Из найденных представительных элементарных классификаторов отбираются наиболее информативные, например те, которые достаточно часто встречаются в описаниях прецедентов. Вычисление оценки принадлежности распознаваемого объекта классу K осуществляется на основе проведения корректной процедуры голосования, в которой участвуют все отобранные представительные элементарные классификаторы класса K . Корректность голосования обеспечивает безошибочное распознавание прецедентов.

В настоящей работе предлагается модифицировать этап обучения на основе применения методов поиска частых (часто встречающихся) фрагментов в описаниях прецедентов класса с последующим анализом встречаемости этих фрагментов в описаниях прецедентов из других классов.

Методы поиска частых элементов в данных обычно используются для построения ассоциативных правил. Наиболее востребованной областью их применения является анализ потребительской корзины [5].

Описываются результаты экспериментов на модельных и реальных данных. Приводятся теоретические выводы о сложности обучения классификаторов типа «Кора», базирующиеся на получении асимптотик для типичных значений наиболее важных количественных характеристик обучающей выборки.

2. Основные результаты

Рассматривается задача классификации по прецедентам с признаками $x_1, \dots, x_n$ и непересекающимися классами K_1 , …, K_l . Каждый признак *имеет* ограниченное число допустимых значений, которые кодируются целыми числами.

Пусть M – исследуемое множество объектов и пусть H – набор из r , $r \le n$, различных признаков вида $\{x_{j_1}, \dots, x_{j_r}\}$; $\sigma = (\sigma_1, \dots, \sigma_r)$; σ_i – допустимое значение признака x_{j_i}, $i = 1, 2, \dots, r$. Пара (σ, H) называется *элементарным классификатором* (далее ЭК) ранга r .

ЭК (σ, H) *порождает* ЭК (σ', H') , если $\sigma' \subset \sigma$, $H' \subset H$. Объект S из M , имеющий признаковое описание $(a_1, \dots, a_n)$, *содержит* ЭК (σ, H) , если $a_{j_t} = \sigma_t$ при $t = 1, 2, \dots, r$.

Пусть K – некоторый класс объектов из M , $\overline{K} = \{K_1, \dots, K_l\} \setminus K$. Положим $Q(K)$ и $Q(\overline{K})$ – множество прецедентов соответственно из K и $\overline{K}$; $|Q(K)| = m_1$, $|Q(\overline{K})| = m_2$, $1 \le p \le m_1$.

ЭК (σ, H) – p -*частый* в K , если не менее p объектов из $Q(K)$ содержат (σ, H) . ЭК (σ, H) называется *максимальным* p -частым в K, если (σ, H) – p -частый в K и в K не существует p -частого ЭК, порождающего (σ, H) .

ЭК (σ, H) ранга r называется *правильным* в K , если (σ, H) – r -*частый* в K .

ЭК (σ, H) – *покрытие* для $\overline{K}$, если ни один прецедент из $Q(\overline{K})$ не содержит (σ, H) . ЭК (σ, H) называется *минимальным* покрытием для $\overline{K}$, если (σ, H) – покрытие

для $\overline{K}$ и (σ, H) не порождает никакого другого покрытия для $\overline{K}$.

ЭК (σ, H) – p -*представительный* для $Q(K)$, если (σ, H) – p -частый в $Q(K)$ и (σ, H) – покрытие для $\overline{K}$. ЭК (σ, H) называется *тупиковым* p -представительным для K , если (σ, H) – p -частый в K и (σ, H) – минимальное покрытие для $\overline{K}$.

В классическом варианте обучение классификатора рассматриваемого типа (модель A_1) основано на построении для каждого класса K тупиковых p -представительных ЭК. Для нахождения искомых ЭК перечисляются минимальные покрытия для $\overline{K}$, среди которых отбираются p -частые в K . В представляемой работе предложена и исследована модель классификатора A_2 , которая основана на перечислении максимальных p -частых ЭК класса K и отборе среди них тех, которые являются покрытиями для $\overline{K}$. В экспериментах для поиска частых ЭК описания прецедентов преобразовывались в бинарные вектора, далее применялся алгоритм FP-Growth [5]. На случайных данных показано, что новая модель классификатора (модель A_2) позволяют существенно сократить временные затраты при построении требуемых ЭК (см. рис. 1).

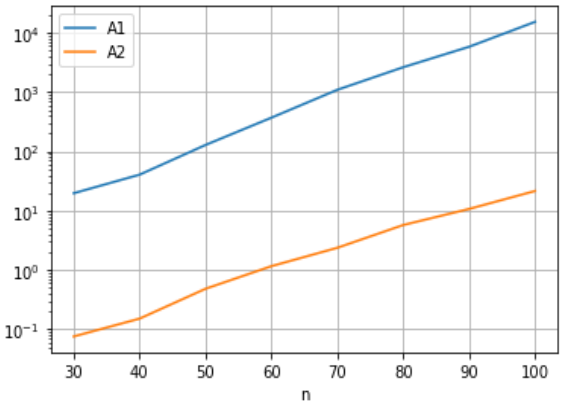

Рис. 1. Зависимость времени работы моделей A_1 и A_2 от n в секундах при $m_1 = 30$, $m_2 = 90$, $p = 1$ (логарифмическая шкала)

Результаты счёта на реальных целочисленных данных приведены в таблице 1 (k – максимальная значность признака). Две первые задачи “Молекулярная биология” и “Шахматы” взяты из репозитория UCI (https://archive.ics.uci.edu/ml/datasets). Третья задача предложена Федеральным медико-биологическим агентством. Модели A_1 и A_2 сравнивались по скорости работы и по точности классификации. В тестировании на точность также участвовал хорошо известный линейный классификатор Логистическая регрессия.

ТАБЛИЦА 1. ВРЕМЯ РАБОТЫ И ТОЧНОСТЬ НА ТЕСТОВОЙ ВЫБОРКЕ

Данные (m_1, m_2, n, k)	A_1 **сек**	A_1 **точн.**	A_2 **сек**	A_2 **точн.**	**Лог. регр.**
Задача 1 $(767, 768, 61, 6)$	3012.2	0.963	**25.9**	**0.967**	0.953
Задача 2 $(1527, 1668, 73, 3)$	323.6	0.951	**180.7**	**0.980**	0.965
Задача 3 $(16, 63, 163, 2)$	370.4	0.750	**89.1**	**0.813**	0.750

Представляет интерес получение асимптотических оценок для типичных значений количественных характеристик обучающей выборки, позволяющих оценить вычислительную сложность исследуемых моделей классификаторов в типичном случае. К таким характеристикам, в частности, относятся число частых ЭК класса K и число минимальных покрытий для $\overline{K}$, а также ранг ЭК указанного вида. В настоящей работе исследован случай бинарных данных.

В случае бинарных признаков множество $Q(K)$ представимо виде булевой матрицы L_K , имеющей размеры $m_1 \times n$. Введем обозначения: $M_{m_1 n}$ – множество всех булевых матриц размера $m_1 \times n$; $R(K)$ – число всех правильных ЭК в K ; $b_n \sim c_n$, $n \to \infty$, означает, что $lim_{n\to\infty} b_n / c_n = 1$; $\phi(m_1)$ – интервал

$$(0.5log_2 m_1 n - 0.5log_2 log_2 m_1 n - log_2 log_2 log_2 n ,$$

$$0.5log_2 m_1 n - 0.5log_2 log_2 m_1 n + log_2 log_2 log_2 n) .$$

Теорема 1. *Если* $m_1{}^a \le n \le 2^{m_1}$, $a > 1$, *то для почти всех матриц* L_K *из* $M_{m_1 n}$ *справедливо*

$$R(K) \sim \sum_{r\in\phi(m_1)} C_n^r C_{m_1}^r 2^{r-r^2} \ , \ n \to \infty ,$$

и ранги почти всех правильных ЭК класса K *принадлежат интервалу* $\phi(m_1)$.

Аналогичные числовые характеристики для ЭК, являющихся минимальными покрытиями для $\overline{K}$, получены в [6]. Сравнение оценок из [6] с оценками, приведёнными в теореме 1, свидетельствует об эффективности в плане вычислительных затрат предлагаемого подхода к обучению классификаторов типа «Кора» в случае, когда m_2 не меньше m_1 , что согласуется с результатами экспериментов.

3. ЗАКЛЮЧЕНИЕ

Исследованы актуальные вопросы снижения временных затрат, возникающие при логическом анализе данных в задачах классификации на основе прецедентов. Теоретически и экспериментально обоснована целесообразность применения методов поиска частых элементов в данных на этапе обучения логического классификатора.

ЛИТЕРАТУРА

[1] Баскакова, Л.В. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств / Л.В. Баскакова, Ю.И. Журавлёв // Ж. вычисл. матем. и матем. физ. – 1981. – Т. 21, № 5. – С. 1264-1275.

[2] Дюкова, Е.В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания / Е.В. Дюкова, Н.В. Песков // Ж. вычисл. матем. и матем. физ. – 2002. – Т. 42, № 5. – С. 711-723.

[3] Дюкова, Е.В. О логическом анализе данных с частичными порядками в задаче классификации по прецедентам / Е.В. Дюкова, Г.О. Масляков, П.А. Прокофьев // Ж. вычисл. матем. матем. физики. – 2019. – Т. 59, № 9. – С. 1605-1616.

[4] Kovshov, N.V. Algorithms for finding logical regularities in pattern recognition / N.V. Kovshov, V.L. Moiseev, V.V. Ryazanov // Computational Mathematics and Mathematical Physics. – 2008. – Vol. 48(2). – P. 314-328.

[5] Aggarwal, Ch.C. Frequent Pattern Mining / Ch.C. Aggarwal, J. Han. – Springer International Publishing, 2014. – 471 p.

[6] Дюкова, Е.В. Задача монотонной дуализации и её обобщения: асимптотические оценки числа решений / Е.В. Дюкова, Ю.И. Журавлёв // Ж. вычисл. матем. и матем. физ. – 2018. – Т. 58, № 12. – С. 2153-2168.