О логической классификации по прецедентам над произведением частичных порядков

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

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

Дюкова А. П.
Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Москва, Россия
anastasia.d.95@gmail.com

***Аннотация*—Рассматриваются вопросы создания алгоритмического обеспечения для одной из центральных задач машинного обучения – задачи классификации по прецедентам. Построены и исследованы оригинальные процедуры логического анализа и классификации целочисленных данных, представимых в виде совокупности элементов декартова произведения конечных частично упорядоченных множеств (произведения частичных порядков). На этапе обучения предлагаемых процедур осуществляется поиск специальных часто встречающихся фрагментов в признаковых описаниях прецедентов, названных правильными представительными элементарными классификаторами. Приведено обоснование эффективности новых распознающих процедур в случае задания линейных порядков на множествах значений признаков.**

Ключевые слова—классификация по прецедентам, корректный логический классификатор, правильный представительный элементарный классификатор, частично упорядоченные данные.

I. Введение

Среди разных подходов к решению задачи классификации по прецедентам важное место занимают методы, основанные на применении аппарата дискретной математики (логические методы анализа данных). Логический подход возник в связи с необходимостью прогнозировать редкие события.

При конструировании логических классификаторов большое внимание уделяется вопросам синтеза корректных алгоритмов, т.е. алгоритмов, не ошибающихся на обучающей выборке. Предполагается, что каждый признак имеет ограниченное множество допустимых значений, которые кодируются целыми числами, и любые два прецедента из разных классов имеют разные описания. Как правило, обучение классификатора сводится к поиску в исходных данных информативных фрагментов описаний прецедентов, называемых представительными элементарными классификаторами (ЭК). Искомые ЭК позволяют различать объекты из разных классов и имеют содержательное описание в терминах той прикладной области, в которой решается задача. По их наличию в описании распознаваемого объекта решается вопрос о его классификации. Однако на этапе обучения возникают сложные в вычислительном плане дискретные задачи. Фундаментальная роль в создании рассматриваемого подхода к задаче классификации принадлежит научным школам чл.-корр. РАН С. В. Яблонского и акад. РАН Ю. И. Журавлева.

Описание логического классификатора может быть дано с использованием терминологии теории логических функций. Тогда представительный ЭК класса K это допустимая конъюнкция (ДК) для не всюду определённой двузначной логической функции F_K , принимающей на целочисленных описаниях прецедентов класса K и других классов соответственно значение 1 и 0. По определению интервал истинности ДК функции F_K имеет непустое пересечение с множеством единиц функции F_K и пустое пересечение с множеством нулей этой функции.

Традиционные схемы логической классификации ориентированы исключительно на случай, когда множество значений каждого признака представляет собой конечную антицепь и для сравнения целочисленных признаковых описаний объектов используется отношение равенства [1 – 4]. Вопросы модификации логических процедур для корректного решения задачи классификации частично упорядоченных целочисленных данных общего вида рассматривались в ряде работ [1, 5, 6]. Особое внимание уделялось случаю, когда множества значений признаков – конечные цепи. С использованием линейных порядков, согласно частоте встречаемости значения признака в классе, были разработаны практические модели логических классификаторов.

В [4] рассмотрена возможность повышения качества и скорости работы логических классификаторов на основе применения методов поиска в описаниях прецедентов каждого класса K часто встречающихся фрагментов специального вида, названных правильными ЭК и с последующим отбором среди них тех ЭК, которые являются ДК для F_K . Исследован случай, когда на множествах значений признаков частичные порядки не заданы. На этапе обучения классификатора, именуемого алгоритмом REC, фактически ищутся ДК функции F_K , каждая из которых, имея ранг r ($r \geq 1$), принимает значение 1 на не менее, чем r прецедентах класса K . Параметр r выбирается с использованием теоретических оценок типичного ранга правильного ЭК. Экспериментально показано, что осуществляемый на этапе обучения поиск искомых ЭК требует меньших временных затрат по сравнению с решением задач, возникающих при реализации классических моделей. Этот вывод подтверждён теоретическими оценками типичного числа правильных ЭК.

В настоящей работе проведена модификация алгоритма REC для работы с частично упорядоченными

данными. Построен новый более эффективный классификатор REC+.

II. Основные результаты

Рассмотрим задачу классификации по прецедентам с множеством целочисленных признаков $\{x_1, \dots, x_n\}$ и непересекающимися классами $\{K_1, \dots, K_l\}$, $l \geq 2$. Пусть исследуемое множество объектов M представимо в виде $M = N_1 \times \dots \times N_n$, где N_j, $j = \overline{1,n}$, – конечное множество допустимых значений признака x_j, на котором задан частичный порядок. Для обозначения того, что b, $b \in N_j$, следует за a, $a \in N_j$, используется запись $a \preccurlyeq b$.

Зададим частичный порядок на множестве M. Пусть $S = (a_1, \dots, a_n)$ и $S^* = (b_1, \dots, b_n)$ – объекты из M, в которых a_j, $j = \overline{1,r}$, и b_j, $j = \overline{1,r}$, – значения признака x_j. Будем считать, что элемент $S^* = (b_1, \dots, b_n)$ *следует* за элементом $S = (a_1, \dots, a_n)$, если $a_j \preccurlyeq b_j$ при $j = \overline{1,r}$.

Элементарным классификатором (ЭК) ранга r назовем пару (σ, H), в которой H – набор из r различных признаков вида $H = \{x_{j_1}, \dots, x_{j_r}\}$, а $\sigma = (\sigma_1, \dots, \sigma_r)$ – набор, в котором σ_i – допустимое значение признака x_{j_i}, $i = \overline{1,r}$. Если объект S из M имеет признаковое описание $(a_1, \dots, a_n)$ и $a_{j_i} \preccurlyeq \sigma_i$ при $i = \overline{1,r}$, то говорят, что ЭК (σ, H) содержится в объекте S. ЭК ранга r называется *правильным для класса* K, $K \in \{K_1, \dots, K_l\}$, если (σ, H) содержится в не менее, чем r прецедентах класса K. ЭК ранга r называется *правильным представительным для класса* K, если (σ, H) – правильный для K и (σ, H) не содержится ни в одном из прецедентов, не принадлежащих классу K.

Предлагаемый классификатор REC+ работает с частично упорядоченными данными по схеме, аналогичной схеме работы классификатора REC. Для каждого класса K ищутся правильные ЭК заданного ранга r. Если на очередном шаге алгоритма найден искомый ЭК, то проверяется его корректность (содержание в прецедентах не из K). Найденные правильные представительные ЭК класса K «голосуют» за отнесение распознаваемого объекта к этому классу.

В экспериментальном исследовании модели REC+ параметр r выбирался с использованием полученных в работе оценок типичного ранга правильного ЭК в предположении, что $N_j = \{0, 1, \dots, k-1\}$, $k \geq 2$, $j = \overline{1,n}$, и элементы в N_j линейно упорядочены в порядке возрастания. Аналогичный результат для случая, когда N_j, $j = \overline{1,n}$, – антицепь приведён в [4].

На реальных задачах проведено экспериментальное сравнение качества работы алгоритмов REC, REC+, Random Forest (RF) и Logistic Regression (LR). Данные взяты из репозитория ФИЦ ИУ РАН и из базы данных UCI (https://archive.ics.uci.edu/). Рассмотрено шесть задач с двумя классами, содержащими m_1 и m_2 прецедентов. Для оценки качества использован хорошо известный функционал – сбалансированная точность. Итоговая оценка качества классификации получена усреднением значения функционала качества по 10 независимым запускам. В каждом запуске исходные данные случайным образом разделялись на обучающую и тестовую выборки в соотношении 4:1. Результаты счета приведены в Таблице I, в которой кроме размера задачи указана средняя значность признака, обозначаемая h.

Таблица I. Качество идентификации

m_1, m_2, n, h	REC	REC+	RF	LR
50, 217, 19, 25	0.570	**0.603**	0.545	0.578
16, 63, 81, 2	0.623	**0.701**	0.542	0.553
38, 107, 35, 10	0.735	0.746	0.742	**0.774**
767, 768, 60, 5	**0.971**	**0.967**	0.960	0.922
1537, 1668, 36, 2	0.974	0.977	**0.988**	0.956
626, 332, 9, 3	0.976	**0.997**	0.939	0.639

Нетрудно видеть, что алгоритм REC+ показывает наилучшее качество среди всех алгоритмов на четырех задачах из шести. На каждой из рассмотренных задач время работы REC не превосходит 1 секунды. Классификатор REC+ превосходит по качеству работы классификатор REC на пяти задачах, однако работает существенно медленнее REC.

III. Заключение

В работе предложена и исследована модификация разработанного в [4] алгоритма голосования по правильным представительным ЭК для случая, когда признаковые описания объектов – элементы декартова произведения конечных цепей. Для рассматриваемого случая получены теоретические оценки типичного ранга правильного ЭК. Экспериментально подтверждена целесообразность задания линейных порядков на множествах значений признаков в соответствии с частотой встречаемости значения признака в классе.

Благодарности

Исследование выполнено за счет гранта Российского научного фонда № 24-21-00301, https://rscf.ru/project/24-21-00301/ .

Литература

[1] 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. и матем. физ. – 2019. – Т. 59, № 9. – С. 1605–1616. DOI: 10.1134/S0044466919090084

[2] Журавлёв, Ю.И. Распознавание. Математические методы. Программная система. Практические применения / Ю.И. Журавлёв , В.В. Рязанов, О.В. Сенько – М.: ФАЗИС, 2006. – 159 с.

[3] Hammer, P.L. Partially defined boolean functions and cause-effect relationships / P.L. Hammer // In: Lecture at the International Conference on Multi-Attrubute Decision Making Via ORBased Expert Systems. – University of Passau, Passau, Germany, 1986.

[4] Dragunov, N.A. Logical Classification Based on Finding Regular Representative Elementary Classifiers / N.A. Dragunov, E.V. Djukova, A.P. Djukova // Journal of Computer and Systems Sciences International. – 2024. – Vol. 63(4). – P. 634–641. DOI: 10.1134/S1064230724700461

[5] Дюкова, Е.В. Корректная классификация по прецедентам: ДСМ-метод над произведением частичных порядков / Е.В. Дюкова, Г.О. Масляков, Д.С. Янаков // Информатика и её применения. – 2024. – Т. 18, Вып. 3. – С. 61–68. DOI: 10.14357/19922264240308.

[6] Anisimova, D.V. Supervised Classification Problem: Searching for Maximum Patterns / D.V. Anisimova, E.V. Djukova, A.P. Djukova // X International Conference on Information Technology and Nanotechnology (ITNT-2024). – Samara, Russian Federation: IEEE Conference proceedings, 2024. – P. 1–4. DOI: 10.1109/ITNT60778.2024.10582366