Классификация по прецедентам: поиск максимальных логических закономерностей

Д.В. Анисимова
Московский государственный университет имени М.В.Ломоносова
Москва, Россия
diana.anisimova.01@mail.ru

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

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

***Аннотация* — Рассматривается логический подход к задаче классификации по прецедентам. Исследуется известная модель классификатора, в которой анализ исходных целочисленных данных сводится к поиску определённых фрагментов в признаковых описаниях прецедентов, называемых максимальными логическими закономерностями. Главный недостаток такой модели заключается в большом числе отказов от классификации. Кроме того, поиск максимальных логических закономерностей – вычислительно сложная дискретная задача. В представляемой работе предложена модификация модели на основе нового более эффективного способа перечисления максимальных логических закономерностей. Экспериментально установлено, что за счёт специального линейного упорядочивания значений признаков можно достичь более высокого качества классификации.**

Ключевые слова — классификация по прецедентам, логический классификатор, максимальная логическая закономерность, представительный элементарный классификатор, декартово произведение частичных порядков, линейный порядок

I. Введение

Настоящая работа посвящена вопросам повышения эффективности методов логического анализа целочисленных данных в задаче классификации по прецедентам. Существует несколько направлений логической классификации целочисленных данных. Одними из основных являются направление LAD (Logical Analysis of Data) [1–3] и направление CVP (Correct Voting Procedures) [4–6]. Оба направления нацелены на поиск определённых фрагментов в признаковых описаниях прецедентов, называемых в LAD и CVP соответственно логическими закономерностями (англ. patterns) и представительными элементарными классификаторами (англ. representative elementary classifiers). Каждый найденный фрагмент позволяет отличать порождающий его прецедент от прецедентов из других классов. Решающее правило основано на анализе встречаемости логических закономерностей в описании распознаваемого объекта.

Рассмотрена известная модель логического классификатора направления LAD, этап обучения которой заключается в поиске так называемых максимальных логических закономерностей классов (англ. maximum pattern) [1, 2]. Нахождение таких закономерностей является сложной в вычислительном плане задачей. Другой существенный недостаток – это отказ от классификации достаточно большого числа прецедентов (модель не корректна).

Большинство алгоритмов поиска максимальных ЛЗ ориентировано на применение методов дискретной оптимизации. В [1] предложен классификатор рассматриваемого вида (далее алгоритм КАР), использующий комбинаторную технику. В представляемой работе на основе модификации алгоритма КАР построен алгоритм КАР1, значительно снижающий время счёта благодаря представлению исходных данных в виде FP-дерева (Frequent Pattern Tree). Конструкция FP-дерева обычно используется для быстрого нахождения в данных (максимальных) частых элементов [7].

В последнее время в направлении CVP появились модели логических классификаторов, ориентированные на решение задач классификации по прецедентам в случае, когда на множествах значений признаков заданы частичные порядки (в этом случае признаковое описание объекта – это элемент декартова произведения конечных частично упорядоченных множеств) [5, 6]. Создание подобных моделей классификаторов обусловлено существованием прикладных задач классификации, качественное решение которых невозможно в рамках традиционной постановки логической классификации.

Авторами настоящей работы разработан и исследован логический классификатор КАР1+, использующий схему работы алгоритма КАР1 и идею обобщения понятия представительного элементарного классификатора на случай частично упорядоченных данных, предложенную в [6]. Проведено тестирование нового классификатора на модельных и реальных задачах с использованием специального линейного упорядочивания значений признаков, встречающихся в описаниях прецедентов. Установлена целесообразность указанного упорядочения значений признаков для повышения точности классификации.

II. Основные понятия и результаты

Задача классификации решается в стандартной для логического подхода постановке, которая заключается в следующем. Рассматривается множество допустимых объектов M, о котором известно, что оно представимо в виде объединения непересекающихся подмножеств (классов) $K_1, \dots, K_l$, $l \geq 2$. В качестве исходной информации задана выборка объектов из M (обучающая выборка). Каждый объект из обучающей выборки (прецедент) представлен набором значений целочисленных признаков $x_1, \dots, x_n$. Для каждого прецедента известен класс, которому принадлежит этот прецедент. Требуется уметь классифицировать объекты множества M , не вошедшие в обучающую выборку.

Введем основные понятия. Будем считать, что признаки бинарные (принимают значения из $\{0,1\}$), иначе применим One-Hot кодирование.

Пусть $K \in \{K_1, \dots, K_l\}$; $R_1(K)$ и $R_2(K)$ – множество прецедентов из K и не из K соответственно; Q – элементарная конъюнкция над переменными $x_1, \dots, x_n$ и N_Q – интервал её истинности. Конъюнкция Q называется *логической закономерностью (ЛЗ)* класса K, если $N_Q \cap R_1(K) \neq \emptyset$ и $N_Q \cap R_2(K) = \emptyset$.

ЛЗ Q класса K называется *максимальной*, если для любой другой ЛЗ $\tilde{Q}$ класса K выполнено $|N_{\tilde{Q}} \cap R_1(K)| \leq |N_Q \cap R_1(K)|$ (здесь $|P|$ мощность множества P).

Пусть $S = (a_1, \dots, a_n)$ – объект из M, $a \in \{0,1\}$, $j \in \{1, \dots, n\}$. Положим $B_j(S,a) = 1$, если $a_j = a$, иначе $B_j(S,a) = 0$;$W_t = 1/|R_t(K)|$, $t \in \{1,2\}$;

$\mu_j^{(t)}(a) = W_t \sum_{S \in R_t(K)} B_j(S,a)$, $t \in \{1,2\}$.

Алгоритм КАР [1] строит каждую максимальную ЛЗ рекурсивно на основе вычисления оценок $\mu_j^{(1)}(a)$ и выделения «пограничных» для класса K прецедентов. На практике оказывается, что найденные максимальные ЛЗ принимают значение 1 на малом числе прецедентов. Это приводит к большому числу отказов от классификации и не гарантирует корректность алгоритма.

Как уже отмечалось во введении, с целью сокращения временных затрат алгоритм КАР1 использует конструкцию FP-дерева для работы с описаниями прецедентов. Для уменьшения числа отказов предлагается запускать алгоритм несколько раз, подавая на вход те прецеденты класса K, на которых ни одна из уже найденных ЛЗ не обращается в 1.

Алгоритм КАР1 модифицирован на случай, когда на множествах значений признаков заданы конечные частичные порядки. Данная модификация названа алгоритмом КАР1+. В КАР1+ при вычислении величины $B_j(S,a)$ отношение равенства заменено отношением предшествования.

В экспериментах применялась быстрая процедура независимого линейного упорядочивания значений признаков, предложенная в [5] и основанная на вычислении для значения $a \in \{0,1\}$ его информативного веса в классе K , равного числу $\mu_j^{(1)}(a) - \mu_j^{(2)}(a)$.

Эксперименты на случайных модельных данных показали, что новые алгоритмы КАР1 и КАР1+ работают в десятки раз быстрее алгоритма КАР.

Результаты счёта на реальных данных приведены в таблице I. Задачи «Машины» (М), «Крестики-нолики» (КН) и «Шахматы» (Ш) взяты из репозитория UCI (https://archive.ics.uci.edu/), а задачи «Инсульты» (И), «Неорганические соединения» (НС) и «Остеогенная саркома» – из репозитория ФИЦ ИУ РАН. Под каждым названием набора данных указано число объектов и число признаков в обучающей выборке. Сравнивались алгоритмы семейства КАР, алгоритм направления CVP (CVP), логистическая регрессия (ЛР), случайный лес (СЛ), градиентный бустинг (ГБ). Нетрудно видеть, что КАР1 и КАР1+ работают гораздо быстрее и точнее КАР. На большинстве задач алгоритм КАР1+ является лидером по точности классификации и на остальных задачах показывает качество, сравнимое с лучшим результатом.

Таблица I. ТОЧНОСТЬ КЛАССИФИКАЦИИ/ ВРЕМЯ РАБОТЫ (В САНТИСЕКУНДАХ)

Алгоритм	**М 1382 6**	**КН 766 9**	**Ш 2556 36**	**И 63 81**	**НС 116 35**	**ОС 213 19**
КАР	0.40/ 602	0.29/ 23205	0.48/ 918630	0.55/ 2053	0.54/ 602530	0.38/ 378726
КАР1	0.94/ 5	0.97/ 19	0.98/ 1362	0.67/ 4	0.55/ 129	0.51/ 16
КАР1+	0.93/ 3	0.97/ 13	**0.99/** 45	**0.70/** 24	**0.83/** 1999	**0.87/** 442
CVP	0.63/ **<1**	**0.98/ <1**	0.97/ 90	0.62/ 14	0.74/ 47	0.57/ **1**
ЛР	0.50/ **<1**	0.58/ **<1**	0.96/ **5**	0.55/ **<1**	0.78/ **1**	0.55/ **1**
СЛ	0.95/ 12	0.92/ 12	**0.99/** 17	0.55/ 9	0.78/ 9	0.52/ 10
ГБ	**0.98/** 10	0.89/ 8	0.98/ 22	0.57/ 5	0.81/ 6	0.54/ 7

III. Заключение

Разработана и исследована новая модель логического классификатора направления LAD, основанная на вычислительно эффективной схеме обучения. Впервые рассмотрены вопросы обобщения классификаторов LAD на случай частично упорядоченных целочисленных данных. Показано, что линейное упорядочение значений признаков по возрастанию их информативных весов позволяет повысить качество классификации.

Благодарности

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

Литература

[1] Ковшов, Н. В. Алгоритмы поиска логических закономерностей в задачах распознавания / Н. В. Ковшов, В. Л. Моисеев, В. В. Рязанов // Ж. вычисл. матем. и матем. физ. – 2008. – Т. 48, №2. – С. 329–344.

[2] Bonates, T. O. Maximum patterns in datasets / T. O. Bonates, P. L. Hammer, A. Kogan // Discrete Appl. Math. – 2008. – Vol. 156(6). – P. 846–861.

[3] Lancia G., Serafini P. Computational Complexity and ILP Models for Pattern Problems in the Logical Analysis of Data // Algorithms. – 2021. – Vol. 14(8). – P. 235.

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

[5] Дюкова, Е. В. О выборе частичных порядков на множествах значений признаков в задаче классификации / Е. В. Дюкова, Г. О. Масляков // Информатика и её применения – 2021. – Т. 15, № 4. – С. 72–78.

[6] 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. 1605–1927.

[7] Charu C. Aggarwal. Frequent Pattern Mining / Charu C. Aggarwal, Jiawei Han // Springer, 2014. – 471 p.