О логической классификации целочисленных данных

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

Г. О. Масляков
Федеральный исследовательский центр «Информатика и управление» Российской академии наук
Москва, Россия
gleb-mas@mail.ru

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

***Аннотация*— Рассматриваются основные подходы к задаче классификации на основе прецедентов, базирующиеся на применении аппарата дискретной математики (логических методов анализа данных). Предлагается общая схема описания логических классификаторов с использованием терминологии процедур корректного голосования.**

Ключевые слова—классификация на основе прецедентов, логический классификатор, отношение частичного порядка, представительный элементарный классификатор, сильная логическая закономерность, ДСМ-метод

1. Введение

Задача классификации на основе прецедентов рассматривается в следующей постановке.

Исследуется некоторое множество объектов M. Известно, что M представимо в виде объединения непересекающихся подмножеств $K_1, \dots, K_l$, называемых классами. Объекты из M описываются признаками $x_1, \dots, x_n$, каждый из которых является некоторой наблюдаемой или измеряемой характеристикой этих объектов и имеет ограниченное число допустимых значений. Значения признаков кодируются целыми числами. Имеется конечный набор $S_1, \dots, S_m$ объектов из множества M, о которых известно, каким классам они принадлежат. Это прецеденты или обучающие объекты. Прецедент S_i, $i \in \{1, \dots, m\}$, задаётся в виде набора $(a_{i1}, \dots, a_{in})$, где a_{ij} – значение признака x_j. Требуется по предъявленному набору значений признаков $(a_1, \dots, a_n)$, описывающему некоторый объект S из M, о котором, вообще говоря, неизвестно, какому классу он принадлежит, определить (распознать) этот класс.

Фундаментальную роль в создании отечественных методов логической классификации сыграли работы члена-корреспондента РАН С.В. Яблонского, в которых введено хорошо известное в дискретной математике понятие теста, и работы академика РАН Ю.И. Журавлева, опубликованные в 70-х и 80-х годах прошлого века. Понятие теста, первоначально применяемое в задачах контроля управляющих систем, явилось основным для конструирования одной из первых моделей классификаторов, именуемых далее процедурами корректного голосования (PCV). Основы проблематики были заложены также в статьях российских ученых М.М. Бонгарда (1967 г.) и М.Н. Вайнцвайга (1973 г.).

В дальнейшем направление PCV развивалось в работах отечественных и зарубежных авторов и существенное развитие получило в статьях [2–6].

Зарубежные исследования в области логической классификации представлены методами Logical Analysis of Data (LAD) и Formal Concept Analysis (FCA). Основополагающие идеи LAD и FCA принадлежат соответственно П. Хаммеру (1986 г.) и Р. Вилле (1981 г.).

В России методы LAD предложены практически параллельно с зарубежными авторами и развиты в ряде работ Ю.И. Журавлёва, В.В. Рязанова (см., например, [7, 9]). Методы FCA представлены в работах В.К. Финна, С.О. Кузнецова, М.И. Забежайло, Д.И. Игнатова и Д.В. Виноградова ([1, 8, 10–12]). В [10] предложен так называемый метод автоматического порождения гипотез (или ДСМ-метод), который позднее в 1990-х годах был адаптирован В.К. Финном и его учениками для задач машинного обучения. ДСМ-классификатор можно отнести к FCA.

Все три названных направления PCV, LAD и FCA имеют много общего. С другой стороны, каждый из подходов использует свою терминологию и демонстрирует некоторую оригинальность. В настоящей работе предлагается общее описание подходов с использованием понятий PCV.

2. Описание процедур PCV, LAD и FCA

Введем основные понятия, используемые при синтезе процедур PCV.

Пусть $H = \{x_{j_1}, \dots, x_{j_r}\}$ – набор из r различных признаков, $\sigma = (\sigma_1, \dots, \sigma_r)$ – набор, в котором σ_i – допустимое значение признака x_{j_i}, $i = 1,2, \dots, r$. Пара (σ, H) называется элементарным классификатором (ЭК) ранга r [6].

Близость объекта $S = (a_1, \dots, a_n)$ из M и ЭК (σ, H), $\sigma = (\sigma_1, \dots, \sigma_r)$, $H = \{x_{j_1}, \dots, x_{j_r}\}$, оценивается величиной $B(S, \sigma, H)$, равной 1, если $a_{j_t} = \sigma_t$ при $t = 1,2, \dots, r$, и равной 0 в противном случае. Если $B(S, \sigma, H) = 1$, то говорят, что объект S содержит ЭК (σ, H).

Множество прецедентов класса K обозначается через $R(K)$. ЭК (σ, H) называется *корректным для класса* K, если для любой пары прецедентов $S \in K$ и $S' \notin K$ не выполнено $B(S, \sigma, H) = B(S', \sigma, H) = 1$. Корректный ЭК (σ, H) класса K называется тупиковым, если любой ЭК (σ', H') такой, что $\sigma' \subset \sigma$, $H' \subset H$, не является корректным для K. ЭК (σ, H) – (*тупиковый*) *представительный* для класса K, если (σ, H) – (тупиковый) корректный ЭК для K и хотя бы один объект из $R(K)$ содержит (σ, H).

При синтезе процедур LAD и в ДСМ-методе используются соответственно понятия «логическая закономерность» (ЛЗ) [7] и «ДСМ-гипотеза» [10, 12].

Представительный для класса K ЭК называется *сильной логической закономерностью*, если он содержится в наибольшем числе прецедентов класса K.

Положим $R_K(\sigma, H) = \{S \in R(K): B(S, \sigma, H) = 1\}$, $|R_K(\sigma, H)|$ – мощность множества $R_K(\sigma, H)$.

Представительный для класса K ЭК (σ, H) *порождает положительную ДСМ-гипотезу для* K, если для любого ЭК (σ', H') такого, что $\sigma \subset \sigma'$, $H \subset H'$, найдётся объект $S \in R_K(\sigma, H)$, не содержащий (σ', H').

Классифицирующий алгоритм A на этапе обучения строит для каждого класса K некоторое множество $P^A(K)$ представительных ЭК. В PCV в качестве элементов множества $P^A(K)$ часто рассматриваются тупиковые представительные ЭК. В LAD строятся сильные логические закономерности, а в ДСМ-методе ЭК, порождающие положительные ДСМ-гипотезы.

В PCV и LAD каждый элемент множества $P^A(K)$ «голосует» за отнесение объекта S классу K. Для оценки принадлежности объекта S классу K суммируются соответственно величины $|R_K(\sigma, H)| \times B(S, \sigma, H)$ и $B(S, \sigma, H)$, $(\sigma, H) \in P^A(K)$.

ДСМ-классификатор действует более строго. Объект S относится к классу K, если S содержит хотя бы один ЭК из $P^A(K)$ и не содержит ни одного ЭК из $P^A(K')$ при $K' \neq K$. В противном случае происходит отказ от классификации.

Предлагаемое описание общей схемы обучения алгоритмов логической классификации основано на приводимых ниже утверждениях 1 – 3.

Пусть $\mathcal{P}(K)$ – множество всех представительных ЭК класса K, на котором задан некоторый частичный (предпорядок) порядок $\preccurlyeq$. ЭК $(\sigma, H) \in \mathcal{P}(K)$ называется *максимальным* относительно частичного (предпорядка) порядка $\preccurlyeq$, если не существует ЭК $(\sigma', H') \in \mathcal{P}(K)$ такого, что $(\sigma, H) \prec (\sigma', H')$.

Зададим на множестве $\mathcal{P}(K)$ отношение частичного порядка $\preccurlyeq_1$. Будем считать, что ЭК $(\sigma_2, H_2) \in \mathcal{P}(K)$ следует за $(\sigma_1, H_1) \in \mathcal{P}(K)$, если $H_2 \subseteq H_1$ и $\sigma_2 \subseteq \sigma_1$ Тогда справедливо

Утверждение 1. *ЭК* (σ, H) *является тупиковым представительным для класса* K *тогда и только тогда, когда* (σ, H) *– максимальный относительно частичного порядка* $\preccurlyeq_1$ *элемент множества* $\mathcal{P}(K)$.

Зададим на множестве $\mathcal{P}(K)$ отношение частичного предпорядка $\preccurlyeq_2$. Будем считать, что ЭК$(\sigma_2, H_2) \in \mathcal{P}(K)$ следует за $(\sigma_1, H_1) \in \mathcal{P}(K)$, если $|R_K(\sigma_1, H_1)| \leq |R_K(\sigma_2, H_2)|$. Тогда справедливо

Утверждение 2. *ЭК* (σ, H) *является сильной ЛЗ класса* K *тогда и только тогда, когда* (σ, H) *– максимальный относительно частичного предпорядка* $\preccurlyeq_2$ *элемент множества* $\mathcal{P}(K)$.

Зададим на множестве $\mathcal{P}(K)$ отношение частичного порядка $\preccurlyeq_3$. Будем считать, что ЭК $(\sigma_2, H_2) \in \mathcal{P}(K)$ следует за $(\sigma_1, H_1) \in \mathcal{P}(K)$, если $R_K(\sigma_1, H_1) \subseteq R_K(\sigma_2, H_2)$ и $H_1 \subseteq H_2$. Тогда справедливо

Утверждение 3. *ЭК* (σ, H) *порождает положительную ДСМ-гипотезу для класса* K *тогда и только тогда, когда* (σ, H) *– максимальный относительно частичного порядка* $\preccurlyeq_3$ *элемент множества* $\mathcal{P}(K)$.

Утверждения 1 – 3 имеют силу и в случае частично упорядоченных целочисленных данных [5].

3. Заключение

В работе описана общая схема синтеза алгоритмов логической классификации. В рамках данной схемы в единой терминологии описаны основные алгоритмы классификации направлений PCV, LAD и FCA. Приведены утверждения показывающие, что каждое из рассмотренных направлений логической классификации ориентировано на задание своего порядка на множестве $\mathcal{P}(K)$ и поиске максимальных относительно заданного порядка элементов.

Литература

[1] Виноградов, Д. В. О представлении объектов битовыми строками для ВКФ-метода / Д. В. Виноградов // Научная и техническая информация, Сер. 2. – 2018. – Т.5. – С. 1–4.

[2] Дюкова, Е. В. Об асимптотически оптимальном алгоритме построения тупиковых тестов / Е. В. Дюкова // Докл. АН СССР, 1977. – Т. 233, №4. – С. 527–530.

[3] Дюкова, Е. В. Дискретный анализ признаковых описаний в задачах распознавания большой размерности / Е. В. Дюкова, Ю. И. Журавлёв // Ж. вычисл. матем. и матем. физ. – 2000. – Т. 40, №8. – С. 1264–1278.

[4] Дюкова, Е. В. Об алгебраическом синтезе корректирующих процедур распознавания на базе элементарных алгоритмов / Е. В. Дюкова, Ю. И. Журавлёв, К. В. Рудаков // Ж. вычисл. матем. и матем. физ. – 1996. – Т. 36, №8. – С. 217–225.

[5] 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, No9. – С. 1605–1616.

[6] 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. 741–753.

[7] Журавлёв, Ю. И. Распознавание. Математические методы. Программная система. Практические применения / Ю. И. Журавлёв, В. В. Рязанов, О. В. Сенько // М.: ФАЗИС. – 2006. – Т. 176. – 159 с.

[8] Забежайло, М. И. О некоторых оценках сложности вычислений в ДСМ-рассуждениях / М. И. Забежайло // Искусственный интеллект и принятие решений. – 2015. – Т. 1. – С. 3–17.

[9] 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.

[10] Финн, В. К. О возможности формализации правдоподобных рассуждений средствами многозначных логик / К. В. Финн // Всесоюзн. симп. по логике и методологии науки. – Киев: Наукова думка, 1976. – С. 82–83.

[11] Gnatyshak, D. V. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns / D. V. Gnatyshak, D. I. Ignatov, S. O. // Kuznetsov Mach Learn. – 2015. – Vol. 101. – P. 271–302.

[12] Kuznetsov, S. O. Mathematical aspects of concept analysis / S. O. Kuznetsov // Journal of Mathematical Science. – 1996. – Vol. 80(2). – P. 1654–1690