

УДК 519.7

КРИТЕРИИ ПОЛНОТЫ ЛИНЕЙНОЙ МОДЕЛИ АЛГОРИТМОВ КЛАССИФИКАЦИИ ОТНОСИТЕЛЬНО РАЗЛИЧНЫХ СЕМЕЙСТВ РЕШАЮЩИХ ПРАВИЛ

© 2020 г. А. Г. Дьяконов^{1,*}, А. М. Головина^{2,**}

Представлено академиком РАН К. В. Рудаковым 20.11.2019 г.

Поступило 29.11.2019 г.

После доработки 29.11.2019 г.

Принято к публикации 29.11.2019 г.

В рамках алгебраического подхода Ю. И. Журавлёва к решению задач классификации рассмотрена линейная модель алгоритмов (оценки принадлежности к классам получаются с помощью линейных регрессий). Исследована возможность ослабить требование полноты (получение произвольной матрицы оценок) для получения произвольных классификаций фиксированной выборки объектов, используя решающие правила специального вида.

Ключевые слова: классификация, линейные алгоритмы, алгебраический подход, корректность, полнота

DOI: 10.31857/S2686954320010129

В работе рассматриваются алгоритмы классификации на l классов [1], которые по контрольной выборке (некоторому множеству q объектов) получают бинарную матрицу классификации размера $q \times l$, ij -й элемент которой — ответ алгоритма на вопрос, принадлежит ли i -й объект j -му классу (1 — да, 0 — нет). Алгоритм классификации выбирается из модели алгоритмов (некоторого параметрического множества), используя обучающую выборку (множество объектов с известной классификацией). Например, параметры алгоритма настраиваются так, чтобы ответ на обучающей выборке был как можно точнее, такая настройка называется обучением.

Следуя алгебраическому подходу к решению задач классификации [2–4], мы сосредоточимся на вопросах реализации произвольных матриц классификации в рамках одной модели, оставляя в стороне классические вопросы машинного обучения (machine learning): проблему выбора параметров, функционалы качества классификации [5] и т.п. В алгебраическом подходе каждый алго-

ритм A представляется в виде суперпозиции $A = B \cdot C$: сначала оператор B формирует матрицу оценок $H = \|h_{ij}\|_{q \times l}$, ij -й элемент которой интерпретируется как оценка принадлежности i -го объекта j -му классу, затем решающее правило C получает по ней матрицу классификации $\|\alpha_{ij}\|_{q \times l}$.

Определение 1. Модель алгоритмов классификации называется *полной* (для данных обучающей и контрольной выборок), если для любой матрицы $H \in \mathbb{R}^{q \times l}$ существует оператор модели B , порождающий матрицу оценок H .

Дальше все определения будут вводиться при фиксированных обучающей и контрольной выборках, поэтому не будем это отдельно обговаривать. Иногда полнота называется *корректностью*, оба понятия были введены в [1] (при неограниченных условиях они эквивалентны). Понятие полноты затем исследовалось в работах [6, 7]. В [8, 9] сделано предположение, что требование полноты слишком жесткое. По сути, полнота нужна для реализации произвольных матриц классификации с помощью достаточно простых решающих правил. Логично потребовать возможность такой реализации и рассматривать нетривиальные решающие правила и/или правила из некоторых семейств.

Определение 2. Модель алгоритмов классификации называется *полной относительно решающего правила C* или *C -полной*, если для любой матрицы $G \in \{0, 1\}^{q \times l}$ суще-

¹Московский государственный университет им. М. В. Ломоносова, Москва, Россия

²Московский государственный технический университет им. Н. Э. Баумана (национальный исследовательский университет), Москва, Россия

*E-mail: djakonov@mail.ru

**E-mail: nastya_gm@mail.ru

стствует такой оператор модели B , что алгоритм $A = B \cdot C$ реализует матрицу классификации G .

О п р е д е л е н и е 3. Модель алгоритмов классификации называется *полной относительно семейства решающих правил C^* или C^* -полной*, если для любой матрицы $G \in \{0, 1\}^{q \times l}$ существуют такие оператор модели B и решающее правило C из семейства C^* , что алгоритм $A = B \cdot C$ реализует матрицу классификации G .

Одно из самых популярных на практике решающих правил [10] — пороговое решающее правило с порогом $c \in \mathbb{R}$:

$$C_c(\|h_{ij}\|_{q \times l}) = \|\alpha_{ij}\|_{q \times l} : \alpha_{ij} = \begin{cases} 1, & h_{ij} \geq c, \\ 0, & h_{ij} < c, \end{cases}$$

для всех $(i, j) \in QL$, здесь и далее $QL = \{1, 2, \dots, q\} \times \{1, 2, \dots, l\}$. Семейство пороговых решающих правил получается варьированием порога:

$$C_* = \{C_c \mid c \in \mathbb{R}\}.$$

О п р е д е л е н и е 4. Модель алгоритмов классификации называется *линейной*, если существуют такие матрицы $H_1, H_2, \dots, H_l$, матрица H_j имеет размер $q \times k_j$, $j \in \{1, 2, \dots, l\}$, что матрица оценок H получается оператором алгоритма модели тогда и только тогда, когда j -й столбец матрицы имеет вид $H_j x_j$, $x_j \in \mathbb{R}^{k_j}$ для всех $j \in \{1, 2, \dots, l\}$.

Линейная модель задается набором матриц $H_1, H_2, \dots, H_l$, мы не будем подробно описывать, как они вычисляются, приведем лишь один пример. Если для каждого класса независимо мы настраиваем линейный регрессор (или логистическую регрессию [5]) на признаковых n -мерных описаниях объектов обучающей выборки (целевой признак здесь равен 1, если объект принадлежит j -му классу, 0 — в противном случае), то получаем линейную (в нашей терминологии) модель, в которой $H_1 = H_2 = \dots = H_l = X$, в матрице X размера $q \times n$ по строкам записаны признаковые описания контрольных объектов. Формально при использовании логистической регрессии модель нелинейная, а именно эта модель наиболее популярная и часто используемая на практике, но из-за использования порогового решающего правила можно привести эквивалентную ей (порождает такие же матрицы классификации) линейную.

В алгебраическом подходе [1, 3] вводятся операции над операторами, они индуцируются операциями над их матрицами оценок. Например, $H(B_1 + B_2) = H(B_1) + H(B_2)$, где $H(B)$ — матрица оценок оператора B . Аналогично вводится умножение на константу. Линейное замыкание модели (говорят сразу про модель алгоритмов, так как

решающее правило фиксируют) — множество алгоритмов, операторы которой являются линейными комбинациями операторов модели. Для линейных моделей линейное замыкание модели совпадает с самой моделью, что освобождает от необходимости отдельно исследовать модель и ее линейное замыкание.

Для матриц $X_1 \in \mathbb{R}^{q \times l_1}$, $X_2 \in \mathbb{R}^{q \times l_2}$ с помощью квадратных скобок будем обозначать их горизонтальную конкатенацию размера $q \times (l_1 + l_2)$:

$$[X_1, X_2].$$

Для большего числа матриц — аналогично, т.е. для линейной модели матрица оценок имеет вид

$$H = [H_1 x_1, H_2 x_2, \dots, H_l x_l],$$

перебором различных $x_1, x_2, \dots, x_l$ получаем все матрицы оценок операторов модели. Через $\tilde{1} = (1, 1, \dots, 1)^T$ будем обозначать вектор, состоящий из одних единиц (размерность будет понятна из контекста), q -мерный вектор по умолчанию рассматривается как матрица размера $q \times 1$, поэтому запись $[H_j, \tilde{1}]$ означает приписывание к матрице H_j единичного столбца.

При доказательстве описанных ниже теорем существенно используется следующий критерий: система неравенств

$$m_i^T x \geq c, \quad i \in I,$$

$$m_i^T x < c, \quad i \notin I,$$

где m_i^T — i -я строка матрицы $M \in \mathbb{R}^{q \times d}$, совместна (относительно неизвестного вектора x) для всех подмножеств $I \subseteq \{1, 2, \dots, q\}$ тогда и только тогда, когда $\text{rank}(M) = q$.

Т е о р е м а 1. *Линейная модель алгоритмов классификации полна относительно решающего правила C_c (порогового с фиксированным порогом c) тогда и только тогда, когда*

$$\forall j \in \{1, 2, \dots, l\} \quad \text{rank}(H_j) = q.$$

Т е о р е м а 2. *Линейная модель алгоритмов классификации полна относительно семейства пороговых решающих правил тогда и только тогда, когда для всех $j \in \{1, 2, \dots, l\}$ выполняется равенство $\text{rank}(H_j) = q$, кроме, быть может, одного, для которого $\text{rank}([H_j, \tilde{1}]) = q$.*

В [8] доказано, что для алгоритмов вычисления оценок (АВО) [1, 3] возможность реализации любой матрицы классификации с помощью порогового решающего правила эквивалентна возможности реализации любой матрицы оценок операторами модели. Для линейной модели и фиксированного порогового решающего правила

это также справедливо, но в случае возможности использовать произвольное пороговое решающее правило линейная модель может порождать любую матрицу классификации и при этом не любую матрицу оценок. Отметим, что на практике часто регрессоры в линейной модели могут получать константные ответы, поэтому

$$\text{rank}(H_j) = \text{rank}([H_j, \tilde{1}]).$$

В задачах классификации с непересекающимися классами каждый объект может принадлежать лишь одному классу, поэтому исследовать надо возможность реализаций матриц классификаций, в каждой строке которых ровно одна единица, а остальные элементы равны нулю, т.е. матриц из множества

$$G_{nn} = \{G \in \{0,1\}^{q \times l} \mid G\tilde{1} = \tilde{1}\}.$$

Обычно используют такое решающее правило:

$$C_{\max}(\|h_{ij}\|) = \|\alpha_{ij}\| : \alpha_{ij} = \begin{cases} 1, & h_{ij} = \max_i h_{it}, \\ 0, & h_{ij} \neq \max_i h_{it}, \end{cases}$$

для всех $(i, j) \in QL$ (если в i -й строке несколько максимальных элементов, то можно считать, что $\alpha_{ij} = 1$ для какого-то одного $j \in \{s \mid h_{is} = \max_i h_{it}\}$).

Определение 5. Модель алгоритмов классификации называется полной относительно решающего правила C в задаче с непересекающимися классами и, если для любой матрицы $G \in G_{nn}$ существует такой оператор модели B , что алгоритм $A = B \cdot C$ реализует матрицу классификации G .

Полнота относительно решающего правила $C_{\max}$ в задаче с непересекающимися классами будем называть $C_{\max}$ -полнотой.

Теорема 3. Линейная модель алгоритмов классификации $C_{\max}$ -полна в задаче с двумя непересекающимися классами тогда и только тогда, когда

$$\text{rank}([H_1, H_2]) = q. \quad (1)$$

Теорема 4. Если линейная модель алгоритмов классификации $C_{\max}$ -полна в задаче с непересекающимися классами из множества $\{1, 2, \dots, l\}$, тогда она $C_{\max}$ -полна в задаче с любым подмножеством классов $J : \emptyset \neq J \subseteq \{1, 2, \dots, l\}$.

В случае произвольного числа классов получаем необходимое условие $C_{\max}$ -полноты:

$$\forall i, j : 1 \leq i < j \leq l \quad \text{rank}([H_i, H_j]) = q,$$

которое не является достаточным.

Если для всех $s \in \{1, 2, \dots, l\}$ выполнено $\text{rank}(H_s) = q$, кроме, быть может, двух $i, j \in \{1, 2, \dots, l\}$, для которых $\text{rank}([H_i, H_j]) = q$, тогда ли-

нейная модель $C_{\max}$ -полна. Полученное достаточное условие $C_{\max}$ -полноты не является необходимым.

Таким образом, в задачах классификации с непересекающимися классами уже для числа классов $l = 2$ критерии $C_{\max}$ -полноты (1) и полноты

$$\text{rank}(H_1) = \text{rank}(H_2) = q$$

различаются, хотя решающее правило $C_{\max}$ является довольно простым и непараметрическим.

ИСТОЧНИКИ ФИНАНСИРОВАНИЯ

Результаты получены в Центре хранения и анализа больших данных МГУ им. М.В. Ломоносова, поддержаны Фондом НТИ (13/1251/2018 от 11.12.2018).

СПИСОК ЛИТЕРАТУРЫ

1. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. 1978. Вып. 33. С. 5–68.
2. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть I // Кибернетика. 1977. № 4. С. 5–17.
3. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть II // Кибернетика. 1977. № 6. С. 21–27.
4. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристических) алгоритмов. Часть III // Кибернетика. 1978. № 2. С. 35–43.
5. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. N.Y.: Springer-Verlag, 2009. 745 p.
6. Журавлёв Ю.И., Рудаков К.В. Об алгебраической коррекции процедур обработки (преобразования) информации // Проблемы прикладной математики и информатики. 1987. С. 187–198.
7. Рудаков К.В. Об алгебраической теории универсальных и локальных ограничений для задач классификации / Распознавание, классификация, прогноз. М.: Наука, 1989. С. 176–201.
8. Дьяконов А.Г. Алгебра над алгоритмами вычисления оценок: монотонные решающие правила // ЖВМиМФ. 2005. Т. 45. № 10. С. 1893–1904.
9. Дьяконов А.Г. Алгебраические замыкания обобщенной модели алгоритмов вычисления оценок // ДАН. 2008. Т. 423. № 4. С. 461–464.
10. Grigorios T., Apostolos P., Weining Q., Stavros V., D'yakonov A., Puurula A., Read J., Svec J., and Semenov S. Wise 2014 challenge: Multi-label classification of print media articles to topics // Lecture Notes in Computer Science. 2014. V. 8787. P. 541–548.

COMPLETENESS CRITERIA FOR A LINEAR MODEL OF CLASSIFICATION ALGORITHMS WITH RESPECT TO FAMILIES OF DECISION RULES

A. G. D'yakonov¹ and A. M. Golovina²

¹Lomonosov Moscow State University, Moscow, Russian Federation

²Bauman Moscow State Technical University, Moscow, Russian Federation

Presented by Academician of the RAS K.V. Rudakov November 20, 2019

Received November 29, 2019

In the framework of Zhuravlev's algebraic approach to classification problems, a linear model of algorithms is investigated (estimates of class membership are generated by linear regressions). The possibility of weakening the completeness requirement (obtaining an arbitrary estimations matrix) to obtain any classifications using special decision rules is investigated.

Keywords: classification, linear algorithms, algebraic approach, correctness, completeness