

АЛГОРИТМ КЛАССИФИКАЦИИ НЕЧЕТКИХ МНОГОМЕРНЫХ ОБЪЕКТОВ

The paper gives a new method of multidimensional fuzzy classification, which can be reduced to solving a linear disjunctive inequality system. The method develops the ideas of authors for realization of a retrieving mechanism based on the tree with the vertices, associated with linear inequalities, directing searching process to demanded final node. The retrieving process consists in transition from one node to another in dependence of fulfillment of the inequalities associated with the nodes. A specificity of the multidimensional fuzzy classification consists in the fact that the suggested algorithm of the retrieving tree building is oriented at minimization of the discrepancy between theoretical and experimental frequencies with which the objects are classified to given clusters. An algorithm for solving a linear disjunctive inequality system is built and theoretically approved. This algorithm is essential for building a retrieving tree. It is based on the iterative process of solving linear inequality systems by means of the residual elimination method.

Введение. Одной из интересных задач в современной рекламной деятельности является задача классификации текстовых документов. Например, при работе с документами в Интернете нередко требуется узнать, какого типа документы просматривал пользователь, чтобы адресовать ему релевантные рекламные документы. Вместе с тем, один и тот же документ может с разной степенью принадлежности ассоциироваться с теми или иными областями деятельности. Возникает задача классификации на множестве нечетких многомерных объектов, если под измерениями текста понимать определенные для него соотношения взаимных частот встречаемости ключевых слов. Предлагаемый в статье механизм использует ассоциированное поисковое дерево.

Задача построения ассоциированного поискового дерева рассматривалась авторами как альтернатива механизму персептрона. При этом задавалась исходная таблица наблюдаемых многомерных объектов, для которых известна их принадлежность к одному из двух множеств (условно, А и В). Требовалось построить ассоциированное бинарное дерево, при движении по узлам которого со значениями каждого из заданных объектов, попадаем в требуемую конечную вершину.

Переформулировка для случая нечетких многомерных объектов выглядит следующим образом. Дано конечное множество векторов n -мерного пространства ($n > 1$) с координатами $x_1, x_2, \dots, x_n$. Для каждого вектора x_k известна нечеткая мера p_k принадлежности его к некоторому (единственному) классу C . Требуется построить линейную различающую функцию

$$F(x) = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n.$$

Для каждого вектора x_k вводим булеву переменную $q_k \in \{0, 1\}$ и два неравенства:

$$\begin{aligned} q_k &\rightarrow a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k \geq 0, \\ a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k &\geq \rightarrow q_k. \end{aligned} \quad (1)$$

Коэффициенты a_i и величины q_k должны определяться из следующего общего условия:

$$\sum_k (q_k - p_k)^2 \rightarrow \min, \quad a_0, a_1, \dots, a_n? \quad (2)$$

Функционал (2) преобразуется к линейной целевой функции после возведения в квадрат выражения под знаком суммы:

$$\begin{aligned} \sum_k (q_k^2 - 2q_k * p_k + p_k^2) &\rightarrow \min \equiv \\ \equiv \sum_k (q_k - 2q_k * p_k + p_k^2) &\rightarrow \min \equiv \\ \equiv \sum_k (1 - 2p_k) * q_k &\rightarrow \min. \end{aligned} \quad (3)$$

Построение системы линейных дизъюнктивных неравенств. Каждое из неравенств (1) приводится к виду дизъюнктивного неравенства:

$$\begin{aligned} q_k &\rightarrow a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k \geq 0 \equiv \\ q_k &= 0 \vee a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k \geq 0 \equiv \\ q_k &\geq 0 \vee q_k \geq 0 \vee a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k \geq 0, \\ a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k &\geq 0 \rightarrow q_k \equiv \\ a_0 + a_1x_1^k + a_2x_2^k + \dots + a_nx_n^k &< 0 \vee q_k = 1 \equiv \\ a_0 + a_1x_1^k + \dots + a_nx_n^k &< 0 \vee q_k \geq 1 - q_k \geq -1. \end{aligned}$$

Таким образом, получена линейная дизъюнктивная система неравенств с целевой линейной функцией и вещественными переменными.

Алгоритм. Сначала рассмотрим алгоритм для случая простых неравенств. Итак, начнем с того, что покажем, как отыскать коэффициенты a_1 и a_2 в случае двумерных векторов. Пусть дана табл. 1:

Таблица 1

x_1	2	-1	0	2	3	6	1	4
x_2	4	3	1	5	-2	3	1	3
y	≥ 0				< 0			

Запишем следующую систему неравенств:

$$\begin{aligned}
 2a_1 + 4a_2 &\geq 0, \\
 -1a_1 + 3a_2 &\geq 0, \\
 0a_1 + 1a_2 &\geq 0, \\
 2a_1 + 5a_2 &\geq 0, \\
 3a_1 - 2a_2 &< 0, \\
 6a_1 + 3a_2 &< 0, \\
 1a_1 + 1a_2 &< 0, \\
 4a_1 + 3a_2 &< 0.
 \end{aligned} \tag{4}$$

Неравенства (4) составлены согласно описанному принципу. Если найти коэффициенты a_1 и a_2 , то все множество записей будет разделено нужным нам образом на две условные половины: «0» и «1». Имеются две проблемы. Во-первых, следует избавиться от жестких неравенств («<»). Это сделать нетрудно, если ввести достаточно малую величину $\xi > 0$, такую что $c_1a_1 + c_2a_2 < 0$ можно заменить на $c_1a_1 + c_2a_2 \leq -\xi$.

Заметим, что такая замена может сделать систему неравенств несовместной, однако и исходная система может быть несовместна, так что наша основная задача – разрешить проблему несовместности системы неравенств. В иллюстративных целях зададим $\xi = 1$ и приведем все неравенства к виду « $\geq$ »:

$$\begin{aligned}
 2a_1 + 4a_2 &\geq 0, \\
 -1a_1 + 3a_2 &\geq 0, \\
 0a_1 + 1a_2 &\geq 0, \\
 2a_1 + 5a_2 &\geq 0, \\
 -3a_1 + 2a_2 &\geq 1, \\
 -6a_1 - 3a_2 &\geq 1, \\
 -1a_1 - 1a_2 &\geq 1, \\
 -4a_1 - 3a_2 &\geq 1.
 \end{aligned} \tag{5}$$

Для решения мы используем стратегию устранения невязок (СУН) [2], которая приводит к решению: $a_1 = -2$, $a_2 = 1$.

Итак, разделяющее неравенство получено в форме

$$-2x_1 + x_2 \geq 0. \tag{6}$$

Пусть нужно провести «разбивку» точек в соответствии с табл. 2.

Таблица 2

x_1	2	-1	0	2	3	6	1	4
x_2	4	3	1	5	-2	3	1	3
y	≥ 0		< 0		≥ 0		< 0	

Снова составляем систему неравенств (для переменных a_3 и a_4).

$$\begin{aligned}
 2a_3 + 4a_4 &\geq 0, \\
 -1a_3 + 3a_4 &\geq 0, \\
 0a_3 - a_4 &\geq 1, \\
 -2a_3 - 5a_4 &\geq 1, \\
 3a_3 - 2a_4 &\geq 0, \\
 6a_3 + 3a_4 &\geq 0, \\
 -a_3 - a_4 &\geq 1, \\
 -4a_3 - 3a_4 &\geq 1.
 \end{aligned} \tag{7}$$

Решая систему неравенств (7), мы столкнулись со следующей ситуацией. Невязки невыполнимы вместе с базисными неравенствами. Следовательно, система противоречива. Это означает, что нельзя подобрать переменные a_3 , a_4 в (7), чтобы разбить нужным образом записи на подмножества «0» и «1». Поступим так: просто исключим невязки, которые с базисными неравенствами дают противоречие.

В итоге не останется ни одной невязки и процесс завершится. Найдем

$$z_1 = z_2 = 0, a_3 = 2, a_4 = -1.$$

и мы получим снова разделяющее неравенство:

$$2x_1 - x_2 \geq 0. \tag{8}$$

Однако одним неравенством (8) нам обойтись не удастся, так как оно неверно разделяет точки, соответствующие исключенным неравенствам. Неравенство (8) разбивает точки так, как показано на рис. 1.

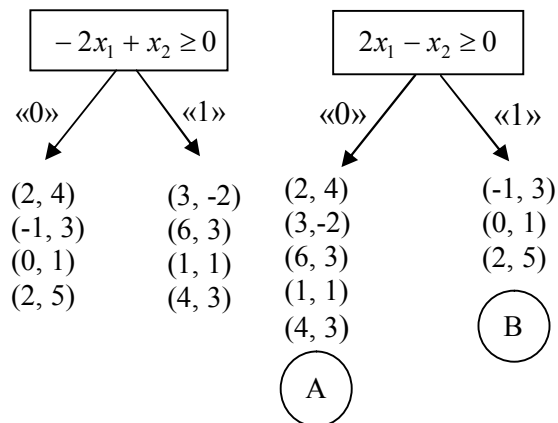


Рис. 1. Разбиение точек на множества

Легко видеть, что ни множество точек на ветви «А», ни множество на ветви «В» не являются «правильными» в смысле рис. 1. Поэтому нам потребуется еще как минимум два неравенства (по одному на каждую из ветвей «А» и «В» на рис. 1), чтобы правильно «развести» точки. Более подробное изложение можно найти в [1].

Решение задач, формализуемых системами линейных дизъюнктивных неравенств, соответ-

вует в целом решению систем простых неравенств и рассматривается для случая, имеющего следующий общий вид:

$$\begin{aligned}
 (1) \quad & x_1 \geq 0; \\
 (2) \quad & x_2 \geq 0; \\
 & \dots \\
 & x_n \geq 0; \\
 & -x_1 \geq -a_1; \\
 & -x_2 \geq -a_2; \\
 & \dots \\
 (2n) \quad & -x_n \geq -a_n, \quad (a_j \geq 0, \quad j = \overline{1, n}); \\
 (2n+1) \quad & (f_1^1 \geq a_{10}^1) \vee \dots \vee (f_{w_1}^1 \geq a_{w_1 0}^1); \\
 (2n+2) \quad & (f_1^2 \geq a_{10}^2) \vee \dots \vee (f_{w_2}^2 \geq a_{w_2 0}^2); \\
 & \dots \\
 (2n+k) \quad & (f_1^k \geq a_{10}^k) \vee \dots \vee (f_{w_k}^k \geq a_{w_k 0}^k).
 \end{aligned} \tag{9}$$

Неравенства с (1) по (2n) системы (9) назовем базовым симплексом.

Каждое l -ое дизъюнктивное неравенство состоит из простых неравенств $f_t^l \geq a_{t0}^l$, где f_t^l означает линейную форму от x_j ($j = \overline{1, n}$), то есть

$$f_t^l \equiv a_{t1}^l x_1 + a_{t2}^l x_2 + \dots + a_{tn}^l x_n.$$

Выполнение дизъюнктивного неравенства соответствует выполнению хотя бы одного входящего в его состав простого неравенства в данной интерпретации I . Невязкой называется такое дизъюнктивное неравенство, в котором каждое простое неравенство представляет невязку. Цель алгоритма, отыскивающего решение (выполняющую интерпретацию) для линейной системы дизъюнктивных неравенств состоит в получении системы без невязок, либо установлении факта невыполнимости системы. Последнее обстоятельство связано с получением в ходе итерации стратегии устранения невязок неравенства (дизъюнктивного или простого), представляющего стоп-невязку. Дизъюнктивное неравенство является стоп-невязкой, если каждое простое неравенство в его составе суть стоп-невязка.

Приведем алгоритм СУН для решения дизъюнктивных неравенств.

Пусть дано некоторое дизъюнктивное неравенство:

$$a_{11}x_1 + \dots + a_{1n}x_n \geq b_1 \vee a_{21}x_1 + \dots + a_{2n}x_n \geq b_2 \vee \dots \vee a_{l1}x_1 + \dots + a_{ln}x_n \geq b_l.$$

Резольвентой этого дизъюнктивного неравенства (в предположении, что $x_j \geq 0, \quad j = \overline{1, n}$) называется простое неравенство вида

$$\max(a_{11}; a_{21}; \dots; a_{l1})x_1 + \dots + \max(a_{1n}; a_{2n}; \dots; a_{ln})x_n \geq \min(b_1; b_2; \dots; b_l).$$

Алгоритм СУН реализуется следующим образом:

1. $i = 1$ (i – номер блочной итерации).
2. На каждой блочной итерации выполнять СУН на матрице, построенной из резольвент текущей системы линейных дизъюнктивных неравенств, дополненной неравенствами вида $x_j \geq 0, \quad j = \overline{1, n}$ до тех пор пока:
 - (А) не будут удалены все невязки;
 - (В) не будет получена стоп-невязка.

Для ускорения сходимости СУН для поиска переменной и невязки, из которой определяется подстановка, применяется градиентный метод. При этом подстановки проводим одновременно как в систему из резольвент, так и в основную систему.

В случае (В) устанавливается невыполнимость системы. В случае (А) – переход на следующий шаг.

3. Если в основной системе нет невязок, то решение найдено, иначе повторение шага 2.

Теорема. Стратегия устранения невязок финитна при решении задачи с дизъюнктивными ограничениями [2].

Заключение. Предложенный алгоритм классификации текстовых документов может быть использован, например, в задаче профилирования пользователей Интернет. Под профилем пользователя понимается множество областей деятельности, к которым он/она имеет интерес. Тогда можно целенаправленно адресовать такому пользователю интересующие его документы, а не забрасывать его почтовый ящик спамом.

Описанный подход допускает сравнительно простой пересчет коэффициентов линейной дискриминаторной функции при добавлении в систему новых данных. Выбор линейной дискриминаторной функции, вообще говоря, влияет только на сложность формируемого поискового дерева, т.к. в силу известного результата Колмогорова-Габора любую полиномиально представимую различающую функцию можно реализовать с помощью линейной функции от большего числа аргументов.

Литература

1. Герман, О. В. Теория информационных процессов и систем. / О. В. Герман, Н. Н. Дорожкина, — Мн.: БГТУ, 2007. — 222 с.
2. Дорожкина, Н. Н. Математические методы и программные средства для исследования и решения задач, формализуемых системами линейных дизъюнктивных неравенств: Автореф. дис... канд. физ.-мат. наук. / Н. Н. Дорожкина, — Ставрополь: СГУ, 2003.