Исследование задачи распределения признаков по классам

Сабурова Мария Ивановна

Аспирант

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: masha-saburova@ya.ru

Рассматривается задача, входные данные которой включают в себя конечную обучающую выборку, в которой каждый объект имеет признаковое описание и метку класса из конечного набора предопределенных классов, включая дополнительный класс. Требуется каждый признак однозначно отнести одному из предопределенных классов.

В ситуациях, когда признаки — это предикаты наличия у объекта некоторого (соответствующего признаку) свойства, принято говорить о маркерах или индикаторах свойств. Тогда рассматриваемую задачу можно воспринимать как задачу однозначной классификации маркеров по классам. Результат решения задачи распределения маркеров по классам интерпретируется как ядра классов и имеет собственную интерпретацию и ценность. Исследуемая постановка задачи востребована в разных предметных областях, например, в области рубрикации текстов [1], в контент-анализе или области анализа экспрессии генов [2].

Для формализации представления данных, в случае, когда признаки— это маркеры, то есть предикаты наличия свойства у объекта, предложена трехдольная модель.

Определение 1. *Трёхдольная модель данных — это реляционная модель, в которой три единицы анализа: объекты, маркеры и классы — и три бинарных гетерогенных отношения между этими единицами анализа.*

Для задачи распределения маркеров по классам формализация в рамках трёхдольной модели получает следующее уточнение — отношение между маркерами и классами является функциональным, то есть это частичное отображение маркеров в исходные классы. Такую трёхдольную модель данных, в которой часть бинарных отношений должна быть частичными отображениями, будем называть *трёхдольной полужесткой*.

Для сведения задачи классификации признаков к задаче классификации объектов была введена линейная информационная модель

с неотрицательными весами признаков, в которую частичная функция из признаков в классы входит явным образом как параметр, т. е. в классификаторе явно происходит приписывание признака к классу.

Пусть T — число признаков, t — номер признака от 1 до T, I — число размеченных объектов, i — номер объекта от 1 до I, J — число классов, j — номер класса от 1 до J, a_t — номер класса, к которому приписан признак t (искомая функция), f_{it} — значение признака t для объекта i (в частности, 0 или 1 для бинарного отношения объект-признак), c_i — истинная метка объекта i.

Модель действует по формуле $\Gamma_k = \sum_{t=1}^{T} w_t f_t[a_t = k]$. Предлагаемый метод обучения модели напоминает многоклассовый SVM [3], и в упрощенном виде в случае линейно разделимой выборки задача обучения выглядит следующим образом:

$$\begin{cases} \dfrac{1}{2}||w||^2 \to \min \\ \displaystyle\sum_t w_t f_{it}([a_t = c_i] - [a_t = j]) \geq 1, \ \forall i, \ \forall j \neq c_i \\ w_i \geq 0, \ \forall t \end{cases}$$

Результаты работы предложенного метода распределения признаков по классам проиллюстрированы на модельных и реальных данных. Предложенная информационная модель и метод её обучения, даже в упрощенном варианте показывают результаты, сравнимые с экспертным мнением.

Исследование выполнено при финансовой поддержке РФФИ (проекты №13-01-00751, №15-07-09214).

Литература

1. Sebastiani F. Machine learning in automated text categorization. ACM computing surveys (CSUR). 2002. Т. 34. №1. С.1-47.
2. Velculescu V. E. et al. Serial analysis of gene expression. Science. 1995. Т. 270. №5235. С.484-487.
3. Duan K. B., Keerthi S. S. Which is the best multiclass SVM method? An empirical study. Multiple Classifier Systems. Springer Berlin Heidelberg. 2005. С.278-285.