Алгебраический подход к задаче иерархической классификации текстов

Остапец Андрей Александрович

Аспирант

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

E-mail: aostapec@mail.ru

Задачу иерархической классификации текстов можно сформулировать следующим образом: имеется множество документов D и множество классов C, которые организованы в иерархию, каждому документу из D приписан один или несколько классов из C. Требуется на основе этих данных построить процедуру автоматической классификации текстов. Структуру классов можно представить в виде дерева, причем классификация проходит только по листьям этого дерева.

Обозначим через $(y_{t1}, \ldots, y_{tl})$ вектор меток: $y_{tj} \in \{0, 1\}$, $y_{tj} = 1 \Leftrightarrow$ t-ая статья принадлежит к классу j. Если $\forall t \in D : \sum_{i=1}^{n} y_{ti} = 1$, то в этом случае с каждой статьей связан *строго* один класс. В некоторых задачах $\exists t \in D : \sum_{i=1}^{n} y_{ti} > 1$, т.е документы могут принадлежать нескольким классам (multi-label classification).

Алгебраический подход [1] заключается в представлении решения задачи в виде суперпозиции двух алгоритмов:

- Первый алгоритм (распознающий оператор) строит вектор оценок принадлежности $(g_{t1}, \ldots, g_{tl})$, где g_{tj} - оценка принадлежности t-й статьи к j-му классу.
- Второй алгоритм (решающее правило) трансформирует вектор оценок $(g_{t1}, \ldots, g_{tl})$ в бинарный вектор $(a_{t1}, \ldots, a_{tl}) \in \{0, 1\}^l$. Ненулевые элементы этого вектора - это классы, к которым относится t-ая статья.

В рамках данной работы в качестве распознающего оператора использовалась линейная комбинация оценок полученных нескольким различными вариантами метода K ближайших соседей (K-Nearest Neighbors). При построении рещающего правила учитывалась специфика задачи и правило строилось в виде: C: $(g_{t1}, \ldots, g_{tl}, \Omega) \to \{0, 1\}^l$, где Ω - это информация об иерархии классов. На открытых данных конкурсов LSHTC1 [2] и LSHTC2 [3] были протестированы различные варианты решающих правил. На этих наборах данных, с помощью представленного подхода, удалось добиться более чем 46% и 37% точности соответственно.

Работа поддержана грантом РФФИ, номер проекта 14-07-00965.

Литература

1. Журавлёв Ю. И. Об алгебраическом подходе. Проблемы кибернетики, 1978, Р. 5–68.
2. Страница конкурса «LSHTC1»: `http://lshtc.iit.demokritos.gr/node/1`
3. Страница конкурса «LSHTC2»: `http://lshtc.iit.demokritos.gr/LSHTC2_CFP`