ИСПОЛЬЗОВАНИЕ НАИВНОГО МЕТОДА БАЙЕСА ДЛЯ КЛАССИФИКАЦИИ КОЛЛЕКЦИИ ТЕКСТОВ

Лой А.А.
Национальный технический университет "Харьковский политехнический институт",
г. Харьков, ул. Пушкинская, 79/2, тел. 707–63–60,
e-mail: loy.alyna@gmail.com

Рост массивов полнотекстовых документов, публикуемых в интернете, требует новых средств организации доступа к информации. Одной из наиболее актуальных проблем управления знаниями, в особенности обеспечения быстрого информационного поиска в полнотекстовых базах знаний, является проблема автоматической классификации набора текстовых документов, которая представляет собой отдельный аспект задачи распознавания смысла текста.

Одним из эффективных алгоритмов классификации является так называемый «наивный» (упрощенный) алгоритм Байеса. Он основан на теореме, утверждающей, что если плотности распределения термов каждого из классов известны, то искомый алгоритм можно выписать в аналитическом виде. «Наивность» алгоритма заключается в предположении, что входные атрибуты условно (для каждого значения класса) независимы друг от друга.

Наивная байесовская модель является вероятностным методом обучения. Следуя предположению, что вероятности попадания термов в определенный класс независимы друг от друга, для получения вероятности в целом достаточно их перемножить. Вероятность того, что документ d попадёт в класс c, записывается как $P(c|d) \propto P(c)\prod_{1\le k\le n_d} P(t_k|c)$.

Здесь $P(t_k|c)$ – условная вероятность того, что терм t_k появится в документе из класса c (оценка вклада термина t_k в то, что документ принадлежит классу c), а $P(c)$ – априорная вероятность того, что документ принадлежит классу c. Последовательность $\langle t_1, t_2, \ldots, t_{n_d} \rangle$ состоит из значащих термов, а n_d – количество таких лексем в документе d. Поскольку цель классификации – найти самый подходящий класс для данного документа, то в наивной байесовской классификации задача состоит в нахождении наиболее вероятного класса c_m

В программной реализации обучающая коллекция представляет собой папку *Tutor*, содержащую набор подпапок, одноименных рассматриваемым классам. Динамически создаются таблицы, соответствующие классам и включающие два поля: индекс терма и его вес в рамках данного класса. Отдельно создаются таблицы *WORDS* (все термы обучающей коллекции) и *CLASSES* (информация о классах). Таблица *WORDS* содержит два поля: сам терм и его индекс; таблица *CLASSES* содержит три поля: название класса, его индекс и количество значащих термов в документах данного класса (без учета индивидуальности).

Чтение документов происходит посимвольно. Термом считается последовательность символов, ограниченных разделителями (все символы, кроме букв латинского алфавита, цифр и ‘&’). Каждый терм проходит морфологическую обработку (отсечение окончания, выявление неправильных глаголов, существительных множественного числа и т.п.). Числительные (последовательности символов, не содержащие букв) и отдельно стоящие буквы отсеиваются до этапа морфологической обработки и не считаются значимыми. Полученный терм сравнивается с таблицей стоп-слов. Стоп-словами считаются союзы, числительные прописью, местоимения, частицы и другие слова, не имеющие семантической значимости (междометия, вводные слова и т.п.). Если совпадений не найдено, терм считается значимым и информация о нем вносится в соответствующие таблицы.

Динамическое создание и заполнение таблиц позволяет реализовать простое редактирование обучающей выборки, а именно: добавление, удаление, объединение и переименование классов; добавление, удаление и редактирование документов, их перемещение в другие классы. Такая организация позволяет включить в интерфейс программы средства редактирования обучающей выборки пользователем.

После обработки документов обучающей коллекции, осуществляется классификация заданной пользователем коллекции. Для этого динамически создаются таблицы для каждого документа, состоящие из двух полей: индекса класса и вероятности $P(c|d)$.

Выделение термов происходит аналогично документам обучающей коллекции, вероятность $P(t_k|c)$ вычисляется для каждого значащего терма на основании информации таблицы *CLASSES* и его веса. К весу каждого терма прибавляется единица, чтобы избежать нулевой вероятности для не встретившихся в определенных классах термов. Термы, отсутствующие в таблице *WORDS* не учитываются. На основании созданной для документа таблицы определяется вероятность принадлежности данного документа к рассматриваемым классам.

Данный способ нашел широкое применение при фильтрации электронной почты, подбора контекстной рекламы, определении области поиска в поисковых системах, решении проблемы омонимии (полисемантичности) слов, что особенно важно для решения задачи автоматизированного перевода.