МЕТОДЫ КЛАССИФИКАЦИИ ТЕКСТОВЫХ ДАННЫХ ПО ТЕМАМ

А.И. Стрелец, *магистр*
В.С. Иванников, *магистр*
А.А. Орлов, *магистр*
А.В. Атавина, *магистр*
Национальный исследовательский ядерный университет «МИФИ»
(Россия, г. Москва)

DOI: 10.24411/2500-1000-2019-11252

Аннотация. *Данная статья описывает методы классификации текстовых данных по темам. Проблема классификации текста является актуальной и важным направлением в области обработки информации и машинного обучения. В статье проведен анализ и исследование существующих решений в данной области. Рассмотрены ключевые моменты различных методов и проведено сих сравнение. На основе проведенного заключения сделаны выводы о специфики применения данных алгоритмов.*

Ключевые слова: *классификация, кластеризация, обработка информации, большие данные, машинное обучение.*

Классификация текстовых данных по темам – это определение принадлежности текстовых данных какой-либо теме, которой посвящён текст. Наиболее часто встречающиеся задачи классификации текстовых данных – это определение эмоциональной окраски текста и классификация текстовых данных по темам. Классификация по темам (многоклассовая классификация) часто используется для фильтрации текстовых данных, когда необходимо отсечь записи, относящиеся к темам, которые не представляют интереса для анализа [1].

Существует несколько подходов к классификации текста. Первый – ручной – представляет собой определение класса документа вручную. Данный метод является точным, но имеет главный недостаток – невозможность обработки больших объёмов данных в приемлемые сроки. Второй подход – написание правил на основе регулярных выражений. Данный подход состоит в том, что специалист по классификации текста составляет набор правил на основе регулярных выражений, что позволяет обрабатывать большие объёмы данных [2]. Однако для создания подобных правил требуются усилия по созданию и поддержанию правил в актуальном состоянии со стороны специалиста. К тому же, перед определением правил специалист должен глубоко ознакомиться с различными образцами данных из всех классов, на что может уйти много времени. Третий подход основывается на машинном обучении. При этом подходе зависимость класса от текста образца определяется автоматически. Данный подход требует предварительной ручной разметки обучающих данных, однако это является более простой задачей, чем определение правил принадлежности всех образцов классам. Такой подход в текущее время является наиболее используемым и перспективным, так как требует наименьшего количества усилий от человека и обладает возможностью автоматической работы с большими объёмами данных.

Классификация текста на основе машинного обучения представляется несколькими основными алгоритмами.

Наивный Байесовский классификатор

Данный метод является методом вероятностной классификации, основанной на теореме Байеса с некоторыми дополнениями. Теорема Байеса даёт отношение между вероятностями двух событий и их условными вероятностями. Наивный Байесовский классификатор предполагает, что наличие или отсутствие определённого свойства класса не имеет отношения к наличию или отсутствию других свойств. Допустим, объект может быть классифи-

цирован по таким атрибутам, как цвет, форма и масса. Резонной классификацией для сферического жёлтого объекта массой менее 60 граммов может быть теннисный мяч. Даже если эти свойства на самом деле зависят друг от друга или являются зависимыми от какого-либо другого свойства, Наивный Байесовский классификатор будет считать, что все эти свойства имеют независимый вклад в то, что объект является теннисным мячом.

Метод опорных векторов

Метод опорных векторов предназначен для решения задач классификации путем поиска хороших решающих границ (рисунок), разделяющих два набора точек, принадлежащих разным категориям. Решающей границей может быть линия или поверхность, разделяющая выборку обучающих данных на пространства, принадлежащие двум категориям.

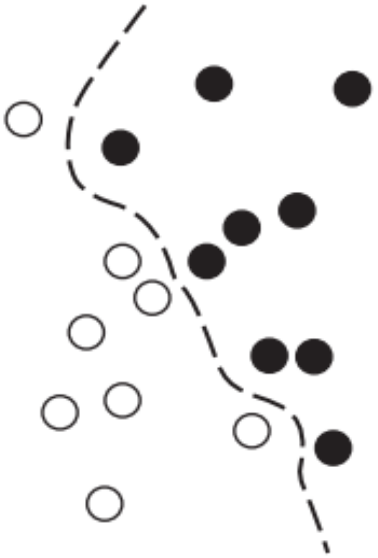

Рисунок. Решающая граница метода опорных векторов

На данный момент метод опорных векторов демонстрировал лучшую производительность на простых задачах классификации. Однако метод опорных векторов оказался трудно применимым к большим наборам данных и не дал хороших результатов для таких задач, как классификация изображений [3]. Так как метод опорных векторов является поверхностным методом, для его применения к задачам распознавания требуется сначала вручную выделить представительную выборку (этот шаг называется конструированием признаков), что сопряжено со сложностями и чревато ошибками.

К плюсам данного метода можно отнести более высокую точность по сравнению с Наивным Байесовским классификатором, а также относительно небольшое количество данных для обеспечения достаточно точных результатов. К минусам – более высокие требования к вычислительным ресурсам (чем Наивный Байесовский классификатор), а также неприменимость к большим наборам данных.

Нейронные сети и глубокое обучение

Нейронная сеть представляет из себя систему из нейронов и связь между ними. В процессе обучение вес связей, соединяющих различные нейроны, постепенно меняет. Результатом обучения нейронной сети является такая сеть, связь которых имеют коэффициент, удовлетворяющие условиям задачи. Нейронные сети характеризуются количеством обучаемых слоёв с нейронными связями. При решении задач с большим объём данных и большим числом параметров, используются нейронные сети с большим числом слоёв.

Обучение таких сетей называется глубоким.

Методы глубокого обучения отличаются двумя важными свойствами: послойное создание сложных представлений, а также детальное исследование промежуточных представлений, благодаря чему каждый слой обновляется в соответствии с потребностями представления слоя выше и потребностями слоя ниже. Вместе эти два свойства делают глубокое обучение намного успешнее предыдущих подходов к машинному обучению [4].

Плюсами нейронных сетей, в том числе сетей глубокого обучения, как правило, является высокая точность классификации текста относительно других методов. К тому же, точность данного метода обычно сильно повышается с увеличением количества данных, т.е. в системах с возможностью накопления данных этот метод будет более подходящим. Минусом нейронных сетей традиционно считается более высокая потребность в вычислительных ресурсах, что раньше тормозило применение нейронных сетей. Однако с распространением вычислений на GPU нейросети стали применяться всё чаще, так как вычисления с их помощью хорошо поддаются распараллеливанию.

Заключение

В результате сравнения существующих методов классификации текстовых данных, можно прийти к следующему заключению. Наиболее подходящим методом для классификации небольшого объёма данных является метод опорных векторов. При обработке большого объёма данных рекомендуется использовать многослойные нейронные сети с глубоким обучением. Метод Байесовского классификатора может использоваться в качестве альтернативного метода, при условии наличия дополнительной информации о разметке данных.

Библиографический список

1. *Морфологический ан*ализатор pymorphy2. – [Электронный ресурс]. – Режим доступа: https://pymorphy2.readthedocs.io/en/latest/ (Дата обращения: 12.04.2019).

2. *Srivastava N., Hinton* G., Krizhevsky A., Sutskever I. and Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting // Journal of Machine Learning Research, pp. 1929-1958, Jun. 2014

3. *Чернодуб А.Н., Дзюба* Д.А. Обзор методов нейроуправления // Проблемы программирования. – 2011. – № 2. – С. 79–94.

4. *Do CIFAR-10 Classifiers* Generalize to CIFAR-10? – [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1806.00451.pdf (Дата обращения: 08.03.2019).

METHODS OF TEXT CLASSIFICATION

A.I. Sagittarius, *master*
V.S. Ivannikov, *master*
A.A. Orlov, *master*
A.V. Atavina, *master*
National research nuclear university "MEPI"
(Russia, Moscow)

Abstract. *This article is about methods of text classification. Problem of classification is one of the most important and actual issue in machine learning and information processing. The article contains analysis and research of existing algorithms. Article also contains the comparation of algorithms. As the result, conclusions were drawn about the algorithm properties.*

Keywords: *classification, clustering, data processing, big data, machine learning.*