Список литературы / References

1. Биоэлектрическое управление / В.С. Гурфинкель, В.Б. Малкин, М.Л. Цетмин, А.Ю. Шнейдер. М: Наука, 1972. 248 с.
2. Протезирование конечностей. [Электронный ресурс]. Режим доступа: http://www.xda.su/artificiallimbs/ArtificiallimbsUpperlimbprostheses/ (дата обращения: 20.03.2017).
3. *Vujaklija I., Farina D., Aszmann O*. New developments in prosthetic leg systems. Orthopedic Research and Reviews, 2016; 8: 31-39.
4. Datasheet. [Электронный ресурс]. Режим доступа: www.alldatasheet.com/datasheet/ (дата обращения: 20.05.2017).
5. Datasheet. [Электронный ресурс]. Режим доступа: www.analog.com/media/en/technical-documentation/data-sheets/ (дата обращения: 20.05.2017).

КЛАССИФИКАЦИЯ ТЕКСТОВ ПРИ ПОМОЩИ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ

Носков Д.В. Email: Noskov640@scientifictext.ru

Носков Дмитрий Владимирович – студент,
департамент информационных технологий и автоматики,
Институт радиоэлектроники и информационных технологий
Уральский федеральный университет им. первого Президента России Б.Н. Ельцина,
г. Екатеринбург

Аннотация: *классификация – одна из основных областей обработки и анализа текстов на естественном языке. В нее входят решения таких задач, как определение тематической принадлежности, определение тональности текста и т.д. Эта область набирает всё большую популярность с каждым годом. Стремительное увеличение объема данных вокруг людей приводит к необходимости разработки эффективных алгоритмов анализа и классификации текстов.*

На сегодняшний день одним из самых распространенных подходов к классификации является подход на основе алгоритмов машинного обучения. Данная статья представляет собой обзор наиболее популярных алгоритмов для построения классификаторов.

Ключевые слова: *классификация текстов, машинное обучение, анализ текстов.*

CLASSIFICATION OF TEXTS USING MACHINE-LEARNING ALGORITHMS

Noskov D.V.

Noskov Dmitrii Vladimirovich – Student,
DEPARTMENT OF INFORMATION TECHNOLOGIES AND AUTOMATION,
INSTITUTE OF RADIOELECTRONICS AND INFORMATION TECHNOLOGIES
URAL FEDERAL UNIVERSITY NAMED AFTER THE FIRST PRESIDENT OF RUSSIA B.N. YELTSIN, EKATERINBURG

Abstract: *classification is one of the main areas of processing and analysis of texts in natural language. It includes solutions to tasks such as determining the subject matter, determining the key of the text, and so on. This area is gaining increasing popularity every*

year. The rapid increase in the volume of data around people leads to the need to develop effective algorithms for analyzing and classifying texts.
To date, one of the most common approaches to classification is the approach based on machine learning algorithms. This article is an overview of the most popular algorithms for building classifiers.
Keywords: *classification of texts, machine learning, text analysis.*

УДК 004.048

Классификация текстов является одной из основных задач компьютерной лингвистики, так как она включает в себя ряд других фундаментальных задач, например, определение тематики или семантический анализ (определение тональности текста).

Использование алгоритмов машинного обучения для решения данных задач достаточно распространенное явление на сегодняшний день, поскольку программы, основанные на данных алгоритмах, имеют достаточно высокий показатель эффективности в сравнении с другими подходами классификации. Обзор и сравнение алгоритмов классификации является достаточно сложной и комплексной задачей, поскольку различные входные данные могут давать разный результат. Поэтому программные реализации алгоритмов необходимо обучать и тестировать на одинаковых наборах данных.

Метод Байеса

Метод основанный на принципе максимума апостериорной вероятности [1].

Пусть $P(c_i|d)$ – вероятность того, что документ d, относится к категории c_i. Задача классификатора заключается в том, чтобы найти такие значения c_i и d, что значение P будет максимальным. Для вычисления P используется теорема Байеса:

$$P(c_i|d) = \frac{P(c_i)P(d|c_i)}{P(d)} \quad (1)$$

где $P(c_i)$ – априорная вероятность, что документ относится к категории c_i; $P(d|c_i)$ – вероятность найти документ d в категории c_i; $P(d)$ – вероятность того, что документ можно представить в виде вектора признаков.

Преимущества: простая программная реализация алгоритма, большая скорость работы;

Недостатки: низкое качество классификации;

Решающее дерево

Дерево принятия решений (дерево классификации, регрессионное дерево) — средство поддержки принятия решений, использующееся в статистике и анализе данных для прогнозных моделей [2]. Структура дерева представляет собой «листья» и «ветки». На ребрах («ветках») дерева решения записаны атрибуты, от которых зависит целевая функция, в «листьях» записаны значения целевой функции, а в остальных узлах — атрибуты, по которым различаются случаи. Чтобы классифицировать новый случай, надо спуститься по дереву до листа и выдать соответствующее значение. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной на основе нескольких переменных на входе. Каждый лист представляет собой значение целевой переменной, измененной в ходе движения от корня по листу. Каждый внутренний узел соответствует одной из входных переменных. Дерево может быть также «изучено» разделением исходных наборов переменных на подмножества, основанные на тестировании значений атрибутов. Это процесс, который повторяется на каждом из полученных подмножеств. Рекурсия завершается тогда, когда подмножество в узле имеет те же значения целевой переменной, таким образом, оно не добавляет ценности для предсказаний.

Преимущества: простота в интерпретации, не требуется подготовка данных, позволяет работать с большим объемом информации без подготовительных процедур;

Недостатки: проблема получения оптимального дерева решений;

Метод опорных векторов

Линейный метод классификации. Для понимания данного метода представим набор документов в виде точек в пространстве размерности |D|. Если точки, принадлежащие разным классам, можно разделить с помощью гиперплоскости (в двумерном случае это прямая), то такую выборку называют линейно разделимой. Очевидно, что для решения данной задачи необходимо провести гиперплоскость так, чтобы точки одного класса лежали по одну сторону, а точки другого класса по другую. Тогда для определения классов неизвестных точек необходимо будет просто посмотреть, с какой стороны от гиперплоскости они находятся. В общем случае можно провести бесконечное множество гиперплоскостей.

На практике редко удается построить гиперплоскость, однозначно разделяющую набор данных. В наборе данных могут иметься такие документы, которые классификатор отнес к одной категории, хотя они должны принадлежать к противоположной. Такие данные создают погрешность в методе опорных векторов.

Метод k ближайших соседей

Метрический метод классификации. Чтобы найти класс, к которому относится документ d, алгоритм сравнивает данный документ со всеми остальными документами обучающей выборки, то есть для каждого d_z вычисляется расстояние $p(d_z, d)$. После этого из обучающей выборки выбираются документы, ближайшие к d. Согласно методу, документ d принадлежит к той категории, которая наиболее распространена среди соседей данного документа.

Преимущества: устойчивость к аномальным выбросам в исходных данных, простая программная реализация, обновление обучающей выборки без переобучения классификатора;

Недостатки: невозможность решения задач большой размерности.

Список литературы / References

1. Байесовский классификатор. [Электронный ресурс], 18 октября 2008. Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Байесовский_классификатор/ (дата обращения: 19.04.2018).
2. Решающее дерево. [Электронный ресурс], 30 января 2010. Режим доступа: http://www.machinelearning.ru/wiki/index.php?title=Решающее_дерево/ (дата обращения: 19.04.2018).