Инженерный вестник Дона, №12 (2022)
ivdon.ru/ru/magazine/archive/n12y2022/8043

Обзор методов классификации текстов с помощью машинного обучения

П.А. Максютин, С.Н. Шульженко

Технологический университет им. А.А. Леонова

Аннотация: Классификация играет большую роль в современном мире. Классификация текстов применяется в решении множества задач, таких, как: распознавание эмоциональной окраски содержимого, определение тематической принадлежности, содержания. В статье рассматриваются различные методы классификации текстов с помощью машинного обучения, их достоинства и недостатки.
Ключевые слова: классификация текстов, машинное обучение, метод опорных векторов, метод Байеса, метод ближайших соседей.

Задача классификации в машинном обучении заключается в определении класса объекта из заранее известных [1]. Классификация текстов состоит из нескольких этапов. Сначала текст необходимо привести в математический вид, удобный для работы классификатора: убрать связочные слова, частицы, междометия, предлоги, провести морфологический анализ [2]. Это значительно уменьшит размерность пространства признаков. Наиболее известными моделями математического представления текста являются модели: «bag of words», «word2vec» и «n-gram».

Модель «bag of words» — это способ представления текста как неупорядоченного набора слов без учета их контекстных связей. Результатом работы данной модели является набор уникальных слов в тексте и их количество повторений. Модель игнорирует контекстные связи слов и также может искажать или терять информацию в словах с несколькими значениями или словах синонимах. Данные проблемы возможно решить путем обогащения модели контекстными связями [3]. Для улучшения масштабирования модели распространен метод хеширования признаков.

Класс моделей «word2vec» предназначен для получения векторных представлений слов. Данный класс моделей основывается на численном представлении слов, которые сохраняют семантическую связь (контекстную

близость): слова, часто встречающиеся в тексте рядом с одинаковыми словами, будут иметь близкие (по косинусному расстоянию) векторы.

Выделяют две модели обучения: CBOW («Continuous Bag of Words») и «Skip-gram» [4,5]. Задача модели CBOW - подобрать слово, основываясь на окружающие слова. Модель «Skip-gram» подбирает окружающие слова, основываясь на входном слове.

В процессе создания модели классификатора, определение значимости признака в тексте является не менее важной задачей, чем построение математического представления текста. Уменьшение размерности пространства признаков может значительно увеличить эффективность работы классификатора и снизить риск переобучения [6]. Выделяют несколько методов определения значимости признака. Наиболее распространенные: LSA, PMI, TF-IDF.

В качестве моделей классификации с помощью машинного обучения можно выделить: KNN, SVM, DT и NB модели.

К ближайших соседей

К ближайших соседей (KNN) - это один из самых простых алгоритмов обучения моделей классификации. Данный метод основан на предположении о том, что похожие объекты расположены близко друг к другу. При этом, для большой размерности пространства признаков, в качестве веса при голосовании можно учитывать расстояние $d(t, t_i)$ от объекта t до каждого из объектов t_i, класс которых известен. В качестве функции расстояния обычно используют евклидову метрику:

$$d(t, t_i) = \sqrt{\sum_{k=1}^{n} (t' - t_i')^2}$$

где n - число характеристик объекта.

При реализации KNN-алгоритма может применять как невзвешенное голосование, так и взвешенное голосование.

При невзвешенном голосовании расстояние от объекта t до каждого из k ближайших соседей t_i не учитывается. Тогда каждый из ближайших соседей t_i голосует за его принадлежность к своему классу $y_{t_i,t}$. Таким образом объект t будет отнесен к классу, который имеет большее кол-во соседей:

$$a = argmax \sum_{i=1}^{n} |y_{t_i,t} = y|$$

При использовании взвешенного голосования учитывается расстояние от объекта t до каждого из ближайших соседей t_i:

$$a = \sum_{i=1}^{k} \frac{1}{d^2(t,t_i)}$$

Главный недостаток метода k ближайших соседей - трудоемкость работы при больших объемах данных и вычисление расстояния до всех объектов при классификации.

Метод опорных векторов

Метод опорных векторов относится к семейству линейных классификаторов. Каждый объект представляется, как точка в n – мерном пространстве. Пространство с меньшей на единицу размерностью, чем объемлющее, называется гиперплоскость. Так, в R^2 пространстве гиперплоскость – это линия. Гиперплоскость называется разделяющей, если сумма расстояний 2-ух ближайших к ней точек с разных сторон максимальна. Метод опорных векторов заключается в поиске разделяющей гиперплоскости. Вектора, проведенные через данные точки, называются опорными.

Гиперплоскость может быть описана уравнением $w * x + b = 0$. Объекты, для которых функция $F(x) = w * x + b = 1$ попадают в один класс, $F(x) = -1$ – в другой. Нахождение разделяющей гиперплоскости заключается в нахождении таких параметров w и b, при которых расстояние до каждого класса объектов максимально. Задача нахождения максимума $\frac{1}{||w||}$ эквивалентна задаче нахождения минимума $||w||^2$:

$$\begin{cases} ||w||^2 \to min \\ y_i(w * x_i + b) \geq 1, i = 1, \dots, m \end{cases}$$

Неразделимой называется такая выборка, в которой точки, относящиеся к разным классам, нельзя разделить гиперплоскостью. В таких случаях необходимо перейти от исходного пространства признаков к новому, в котором обучающая выборка окажется линейно разделимой. Для этого все элементы обучающей выборки преобразуются в более высокую размерность с помощью отображения $\varphi: R^n \to X$. При этом классифицирующая функция примет вид $F(x) = w * \varphi(x) + b$. В случае, когда выборка неразделима, задачу поиска разделяющей гиперплоскости можно свести к задаче поиска седловой точки функции Лагранжа.

Преимущество метода опорных векторов заключается в возможности работы с небольшим набором обучающей выборки и возможности сведения задачи к задаче выпуклой оптимизации, имеющей единственное решение. Стоит так же отметить эффективность применения двухэтапного метода классификации с применением kNN и SVM алгоритмов [7].

Деревья решений

Дерево решения – это ациклический граф, который построен на базе решающих правил, в соответствии с которым делается классификация объектов, описанных набором признаков. В листьях дерева находятся значения целевой функции, в узлах – условия перехода.

Очевидно, чтобы попасть в тот или иной узел, объект должен удовлетворять всем условиям перехода, лежащим на пути к данному узлу. Поскольку путь в дереве к каждому узлу единственный, то и каждый объект может попасть только в один узел. Для борьбы с переобучением модели применяют несколько подходов: отсечение ветвей и ограничение глубины дерева.

К преимуществам данного метода можно отнести простоту реализации и интерпретации результатов. К недостаткам - сложность получения оптимального дерева решений и необходимость иметь большой набор тренировочного множества для точного обучения дерева.

Метод Байеса

Метод Байеса (Naive Bayes) – относится к классу вероятностных алгоритмов классификации. Данный метод позволяет определить класс объекта, основываясь на априорном (предшествующем) распределении вероятности классов данного объекта. Объекту присваивается класс с наибольшей апостериорной вероятностью [8].

Пусть задано множество объектов, каждый из которых представлен вектором признаков $\mathrm{x} = (\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_\mathrm{n})$. Классификатор присваивает каждому объекту условную вероятность $p(C_k|x_1, x_2, \ldots, x_n)$, где C_k - класс.

Пусть $\mathrm{P}(\mathrm{c}_\mathrm{i}, \mathrm{x})$ - вероятность, что объект соответствует классу c_i. Задача классификации методом Байеса заключается в подборе таких значений c_i и x, при которых значение вероятности $\mathrm{P}(\mathrm{c}_\mathrm{i}, \mathrm{x})$ будет максимальным:

$$a = argmaxP(c_i|x)$$

Для вычисления значения $P(c_i, x)$ используется теорема Байеса:

$$P(c_i|x) = \frac{P(x|c_i)P(c_i)}{P(x)}$$

где $P(c_i|x)$ - апостериорная вероятность, $P(x|c_i)$ - вероятность признака с учетом класса, $P(c_i)$ - вероятность класса, $P(x)$ - вероятность признака.

При условии, что признаки независимы, можно воспользоваться формулой:

$$P(d|c_i) = \prod_{k=1}^{n} P(t_k|c_i)$$

Тогда классификатор можно представить, как функцию:

$$y = argmax \prod_{k=1}^{n} P(x_i|C_k)$$

Преимущества данного метода: Простота реализации и интерпретации результатов. Недостатки: Низкое качество классификации. Комбинирование данного метода с другими частично решает данную проблему.

Заключение

Задачи классификации текста являются актуальными в связи с огромным объемом данных, создаваемым пользователями в электронном виде.

В данной статье были рассмотрены методы представления текста в математическом виде: «bag of words», «word2vec» и методы обучения моделей классификаторов: KNN, SVM, DT, NB. Существует множество методов классификации, которые используют различные подходы. Тем не менее эффективность методов зависит от конкретной задачи и области применения. Так,

чтобы обойти недостатки отдельных алгоритмов, набирает популярность подход комбинирования методов классификации [9,10].

Литература

1. Петровский М.И., Глазкова В.В. Алгоритмы машинного обучения для задачи анализа и рубрикации электронных документов. // Вычислительные методы и программирование. 2007. Т. 8. № 2. С. 57-69.

2. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных. М.: НИУ ВШЭ, 2017. 269 с.

3. Нугуманова А.Б., Бессмертный И.А., Пецина П., Байбурин Е.М. Обогащение модели Bag-of-words семантическими связями для повышения качества классификации текстов предметной области // Программные продукты и системы. 2016. №2. С. 89-99.

4. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. // 2013. URL: arxiv.org/abs/1301.3781v3.

5. Mikolov T, Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality. // 2013. URL: arxiv.org/abs/1310.4546.

6. Максаков А. В. Исследование способов уменьшения набора характеристик в алгоритмах классификации текстов. Четвертый российский семинар РОМИП. С. 92-100.

7. Демидова Л. А., Соколова Ю. С. Классификация данных на основе SVM-алгоритма и алгоритма k-ближайших соседей // Вестник РГРТУ. 2017. №62. URL: vestnik.rsreu.ru/ru/archive/2017/4-vypusk-62/516-1995-4565-2017-62-4-119-132.

8. Котельников Е.В., Клековкина М.В. Автоматический Анализ Тональности Текстов На Основе Методов Машинного Обучения // труды XVIII Международной конференции "Диалог 2012": в 2-х томах. Том Выпуск

11 (18), Том 2. Российский государственный гуманитарный университет. М.: Компьютерная Лингвистика И Интеллектуальные Технологии, 2012. С. 27-36.

9. Красников И.А., Никуличев Н.Н. Гибридный алгоритм классификации текстовых документов на основе анализа внутренней связности текста // Инженерный вестник Дона. 2013. №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

10. Поляков И.В., Соколова Т.В., Чеповский А.А., Чеповский А.М. Проблема классификации текстов и дифференцирующие признаки // Вестн. НГУ. Сер.: Информационные технологии. 2015. Т. 13. Вып. 2. С. 55–63.

References

1. Petrovskiy M.I., Glazkova V.V. Vychislitel'nyye metody i programmirovaniye. 2007. T. 8. № 2. pp. 57-69.

2. Bol'shakova YE.I., Vorontsov K.V., Yefremova N.E., Klyshinskiy E.S., Lukashevich N.V., Sapin A.S. Avtomaticheskaya obrabotka tekstov na yestestvennom yazyke i analiz dannykh [Automatic processing of natural language texts and data analysis] M.: NIU VSH·E, 2017. 269 p.

3. Nugumanova A.B., Bessmertnyy I.A., Petsina P., Bayburin YE.M. Programmnyye produkty i sistemy. 2016. №2. pp. 89-99.

4. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. URL: arxiv.org/abs/1301.3781v3.

5. Mikolov T, Sutskever I., Chen K., Corrado G., Dean J. Distributed Representations of Words and Phrases and their Compositionality. 2013. URL: arxiv.org/abs/1310.4546.

6. Maksakov A. V. Issledovaniye sposobov umen'sheniya nabora kharakteristik v algoritmakh klassifikatsii tekstov. Chetvertyy rossiyskiy seminar ROMIP. pp. 92-100.

7. Demidova L. A., Sokolova YU. S. Vestnik RGRTU. 2017. №62. URL: vestnik.rsreu.ru/ru/archive/2017/4-vypusk-62/516-1995-4565-2017-62-4-119-132.

8. Kotel'nikov YE.V, Klekovkina M.V. Avtomaticheskiy Analiz Tonal'nosti Tekstov Na Osnove Metodov Mashinnogo Obucheniya trudy XVIII Mezhdunarodnoy konferentsii "Dialog 2012": v 2-kh tomakh. Tom Vypusk 11 (18), Tom 2. Rossiyskiy gosudarstvennyy gumanitarnyy universitet. M.: Komp'yuternaya Lingvistika I Intellektual'n·yye Tekhnologii, 2012. pp. 27-36.

9. Krasnikov I.A., Nikulichev N.N. Inzhenernyj vestnik Dona. 2013. №3. URL: ivdon.ru/ru/magazine/archive/n3y2013/1773.

10. Polyakov I.V., Sokolova T.V., Chepovskiy A.A., Chepovskiy A.M. Vestn. NGU. Ser.: Informatsionn·yye tekhnologii. 2015. T. 13. Vyp. 2. pp. 55–63.