DOI 10.37882/2223-2966.2023.09.09

ИСПОЛЬЗОВАНИЕ ГЕНЕРАТИВНЫХ АЛГОРИТМОВ ДЛЯ ФОРМИРОВАНИЯ ДОКУМЕНТОВ

USING GENERATIVE ALGORITHMS TO GENERATE DOCUMENTS

A. Kasymov
Yu. Maximov

Summary. This article provides a brief overview of the latest text classification models with an emphasis on data flow, from raw text to output labels. The differences between earlier methods and later methods based on deep learning are emphasized, both in their functioning and in how they transform input data. To give a better idea of text classification, an overview of the data sets for the language is provided, as well as instructions for synthesizing two new data sets with multiple labels. At the end, we describe an overview of new experimental results and discuss the problems of open research related to language models based on deep learning.

Keywords: text classification, tokenization, topic labeling, news classification, transformer, surface learning, deep learning, multicomponent corpora.

Касымов Алексей Алексеевич
Аспирант, Воронежский государственный технический университет
kasimlele@live.ru
Максимов Юрий Максимович
Аспирант, Воронежский государственный технический университет
yuramaximo@mail.ru

Аннотация. В данной статье представлен краткий обзор последних моделей классификации текста с акцентом на поток данных, от необработанного текста до выходных меток. Подчеркиваются различия между более ранними методами и более поздними методами, основанными на глубоком обучении, как в их функционировании, так и в том, как они преобразуют входные данные. Чтобы дать лучшее представление о классификации текстов, предоставляется обзор наборов данных для языка, а также инструкции по синтезу двух новых наборов данных с несколькими метками. В конце описан обзор новых экспериментальных результатов и обсуждаем проблемы открытых исследований, связанные с языковыми моделями, основанными на глубоком обучении.

Ключевые слова: классификация текстов, токенизация, маркировка темы, классификация новостей, трансформатор, поверхностное обучение, глубокое обучение, многокомпонентные корпуса.

Введение

Классификация текста (КТ) является задачей фундаментальной важности, и она набирает обороты благодаря последним разработкам в области анализа текста и обработки естественного языка. Методы классификации текста имеют общую цель — назначить предопределенную метку для заданного входного текста, хотя это наименование может относиться к множеству специализированных методов, применяемых к разным предметным областям.

Классические примеры КТ включают поиск информации, маркировку тем, анализ настроений и классификацию новостей. Однако у КТ есть практические приложения, выходящие за рамки простой категоризации, такие как экстрактивные системы ответов на вопросы и обобщения. В этом случае интуитивное понятие «метка» заменяется выбором между кандидатами (например, ответ или предложение для включения в резюме).

Скорость, с которой в настоящее время создается текстовая информация, давно превзошла ручное решение этих задач, а это означает, что методы КТ не только полезны, но и строго необходимы. Соответственно, разработка точных и непредвзятых систем КТ имеет первостепенное значение.

Текстовое представление

Важным шагом, требуемым любой процедурой КТ, является проекция текстовых признаков в выбранном пространстве признаков. Из-за его неструктурированности (с вычислительной точки зрения) необходимо применить ряд операций, чтобы постепенно преобразовать его в удобоваримую для компьютера форму. Предварительная обработка должна учитывать модели, которые предназначены для использования на более поздних этапах конвейера классификации, поскольку универсального решения не существует.

В частности, более ранние методы в значительной степени полагаются на этап ручного проектирования функций, что требует внимательного отношения и знаний в предметной области. С другой стороны, более поздние методы, основанные на глубоком обучении, заметно отличаются из-за автоматического извлечения признаков. Как мы увидим, предварительная обработка попрежнему важна для этих методов, хотя она может применяться по-разному из-за допущений, которые они делают.

Широкая категоризация методов классификации текста

Поверхностные подходы к обучению

Более ранние методы часто определяются как подходы «поверхностного обучения». Однако, поскольку это определение не является особенно стандартизированным или согласованным, мы уточняем, что с этим термином мы имеем в виду все те традиционные или классические методы, связанные с обычным машинным обучением. То есть в эту группу входят все те методы, предшествующие нейронным сетям, предсказание которых основано на ручных функциях. Кроме того, в эту категорию также входят нейронные сети с очень небольшим (0–2) скрытым слоем, которые сами по себе называются «мелкими» и которые ликвидируют разрыв между этой группой методов и их преемниками, основанными на глубоком обучении.

Поверхностные подходы к обучению являются преемниками подходов, основанных на правилах, которые они превзошли как по точности, так и по стабильности. Поверхностные методы обучения по-прежнему популярны во многих практических контекстах или в качестве надежной основы. Хотя они плохо масштабируются для больших объемов данных, они проявляют себя, когда ресурсов слишком мало, чтобы глубокие методы были эффективными. Эти классические подходы требуют этапа разработки функций, который может быть дорогостоящим в зависимости от сложности предметной области. В то время как вычислительная сторона этих затрат может быть значительной, требования к знанию предметной области, которые необходимы для правильного применения соответствующих методов извлечения признаков, могут быть более трудными для выполнения на практике.

Подходы к глубокому обучению

Появление моделей глубокого обучения затронуло все области искусственного интеллекта, включая классификацию текстов. Эти методы получили распространение из-за их способности моделировать сложные объекты без необходимости их ручной разработки, что устраняет часть требований к знанию предметной области. Вместо этого работа была направлена на разработку архитектур нейронных сетей, способных извлекать эффективные представления для текстовых единиц. Недавние разработки были особенно успешными в этом, породив семантически значимые и контекстуальные репрезентации. Автоматическое извлечение признаков особенно полезно при моделировании текстовых данных, поскольку оно способно использовать базовую лингвистическую структуру документа. Эта структура интуитивно понятна нам, если мы понимаем язык, но обычно непонятна машине.

Основные отличия и вклады

В недавних публикациях методы классификации текстов исследовались с общей точки зрения. Среди них отметим работу Li et al. [1], которая обеспечивает полное исследование моделей, начиная от мелких и заканчивая глубокими. Обзор Kowsari et al. [2] обеспечивает отличное исследование этапов предварительной обработки, таких как извлечение признаков и уменьшение размерности. С другой стороны, работа Minaee et al. [3] сосредоточена исключительно на тщательном изучении глубоких подходов, хотя она также предоставляет количественные результаты для классических методов при анализе экспериментальных характеристик.

Эта работа направлена на то, чтобы обогатить обзоры текстовой классификации, давая обзор каждого шага, связанного с разработкой классификатора текстовых данных. Поэтому мы даем подробное описание наиболее важных операций подготовки данных, используемых совместно с алгоритмами классификации текста. Эти этапы конвейера КТ часто упускают из виду, однако понимание их использования и мотивации их выбора может оказаться основополагающим в построении эффективной основы для этой задачи. Мы продолжаем обобщать информацию об основных наборах данных языка КТ и общий эталон современных подходов в различных подзадачах. Кроме того, мы предоставляем результаты по двум недавно синтезированным наборам данных КТ с несколькими метками, излагая процесс их воспроизведения. Мы считаем, что это важный вклад, поскольку подзадачи, которые они решают, представлены недостаточно.

Предварительная обработка

Входные данные для задач на естественном языке, таких как КТ, состоят из необработанного неструктурированного текста. Текстовая информация, в отличие от других типов данных, таких как изображения или временные ряды, не имеет собственного числового представления; прежде чем передать его любому классификатору, он должен быть спроецирован в соответствующее пространство признаков. Поэтому процедуры предварительной обработки имеют особое значение, поскольку без них нет основы ни для процедур извлечения признаков, ни для алгоритмов классификации.

Стандартные операции предварительной обработки

Токенизация

Самая основная операция предварительной обработки, которая должна применяться к тексту, — это токенизация. Эта процедура определяет уровень де-

тализации, на котором мы анализируем и генерируем текстовые данные, и в целом может быть описана как процесс разбиения потока текста на более мелкие фрагменты (исторически называемые токенами). До недавнего времени в большинстве моделей НЛП в качестве атомарной единицы выбора использовались слова, но недавние подходы заключались в разложении текста на более мелкие единицы (такие как символьные n-граммы или даже более максимальные формы разложения, такие как базовые байты [4]).

Токенизация: документ обрабатывается как строка, а затем разбивается на список токенов.

Удаление стоп-слов: Стоп-слова, такие как «и», «а», «но» и т. д., встречаются часто, поэтому незначащие слова необходимо удалить.

Основополагающее слово: применение алгоритма основообразования, который преобразует другую форму слова в аналогичную каноническую форму. Этот шаг представляет собой процесс объединения токенов с их корневой формой, например,. соединение для соединения, вычисление для вычисления и т.д. (рис. 1)

Рис. 1. Процесс классификации документов

Набор токенов, созданный процедурой токенизации, может содержать ненужные или вводящие в заблуждение элементы. Текстовый шум, такой как специальные символы или лишние символы, должен быть удален. Может быть полезно удалить стоп-слова [7], т.е. неинформативные слова, встречающиеся в большом количестве, но не несущие семантической значимости. Другие процедуры нормализации, такие как приведение к нижнему регистру, исправление орфографических ошибок и стандартизация сленговых слов и сокращений, могут быть полезны для уменьшения количества различных элементов в пространстве признаков.

Предварительная обработка для глубоких моделей

Выбор функции

После выделения признаков важным шагом в предварительной обработке текстовой классификации является выбор признаков для построения векторного пространства, которое улучшает масштабируемость, эффективность и точность текстового классификатора. В общем, хороший метод выбора признаков должен учитывать характеристики предметной области и алгоритма [15]. Основная идея FS состоит в том, чтобы выбрать подмножество функций из исходных документов. FS выполняется путем сохранения слов с наивысшим баллом в соответствии с заранее определенной мерой важности слова [9]. Выбранные признаки сохраняют исходное физическое значение и обеспечивают лучшее понимание данных и процесса обучения [11]. Для классификации текста основной проблемой является высокая размерность пространства признаков. Почти каждая текстовая область имеет большое количество признаков, большинство из которых не имеют значения и не полезны для задачи классификации текста, и даже некоторые шумовые признаки могут резко снизить точность классификации [6]. Следовательно, FS обычно используется в текстовой классификации для уменьшения размерности пространства признаков и повышения эффективности и точности классификаторов.

Было изучено множество показателей оценки признаков, среди которых не удалось выделить прирост информации (IG), частоту терминов, хи-квадрат, ожидаемую перекрестную энтропию, отношение шансов, вес свидетельства текста, взаимную информацию, индекс Джини. Частота терминов и частота документов (TF/DF) (Таблица 1) и т.д. Хорошая метрика выбора признаков должна учитывать характеристики предметной области и алгоритма.

Методы машинного обучения

Документы можно классифицировать тремя способами: неконтролируемыми, контролируемыми и полуконтролируемыми методами. Недавно было предложено много методов и алгоритмов для кластеризации и классификации электронных документов. В этом разделе основное внимание уделялось контролируемым методам классификации, новым разработкам и освещались некоторые возможности и проблемы с использованием существующей литературы. Автоматическая классификация документов по предопределенным категориям вызвала активное внимание, поскольку уровень использования Интернета быстро увеличился. За последние

Таблица 1.

Методы выбора признака

Коэффициент усиления (GR)	$GR(t_k,c) = \frac{c \in \left\{\sum_{c_i \bar{c}_i}\right\} t \in \left\{\sum_{t_k \bar{t}_k}\right\} p(t,c)\log\frac{P(t,c)}{P(t)P(c)}}{c \in \left\{\sum_{c_i \bar{c}_i}\right\} P(c)\log P(c)}$
Информационное усиление (IG)	$IG(w) = -\sum_{j=1}^{K} P(c_j)\log P(c_j) + P(w)\sum_{j=1}^{K} P(c_j\|w)\log P(c_j\|w)$
Условная взаимная информация (CMI)	$CMI(C\|S) = H(C) - H(C\|S_1,S_2,\ldots,S_n)$
Частота документа (DF)	$DF(t_k) = P(t_k)$
Периодичность (TF)	$tf(f_i,d_j) = \frac{freq_{ij}}{\max_k freq_{Kj}}$
Обратная частота документа (IDF)	$\|idf\| = \log\frac{\|D\|}{\|f_1\|}$
Срок (s)	$s(t) = P(t \in y \| t \in x)$
Нечетное соотношение (OR)	$OR(f_i,c_j) = \log\frac{P(f_i\|c_j)(1 - P(f_i\|\neg c_j))}{(1 - P(f_i\|c_j))(P(f_i\|\neg c_j))}$

несколько лет задача автоматической классификации текста была тщательно изучена, и в этой области наблюдается быстрый прогресс, включая подходы к машинному обучению, такие как байесовский классификатор, дерево решений, K-ближайший сосед (KNN), машины опорных векторов (SVM), нейронные сети, латентный семантический анализ, алгоритм Роккио, нечеткая корреляция и генетические алгоритмы и т.д. Обычно для автоматической классификации текста используются методы обучения с учителем, когда документам присваиваются предварительно определенные метки категорий на основе вероятности, предложенной обучающий набор размеченных документов. Некоторые из этих методов описаны ниже.

Алгоритм А. Роккио

Алгоритм Роккио [7] представляет собой метод векторного пространства для маршрутизации или фильтрации документов при информационном поиске, построения вектора-прототипа для каждого класса с использованием обучающего набора документов, т.е. среднего вектора по всем векторам обучающих документов, принадлежащих классу c_i, и вычисления подобия между тестовым документом и каждым из векторов-прототипов, которые относят тестовый документ к классу с максимальным сходством.

$$C_i = \alpha * centroid_{C_i} - \beta * centroid_{\overline{C_i}} \quad (1)$$

При задании категории вектору документов, принадлежащих к этой категории, присваивается положительный вес, а векторам остальных документов присваивается отрицательный вес. Получены положительно и отрицательно взвешенные векторы, вектор-прототип этой категории (рис. 2).

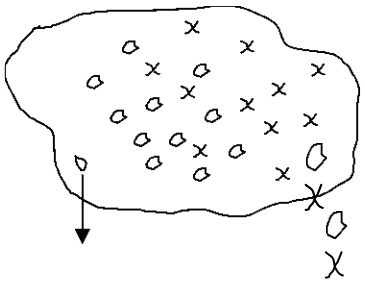

Рис. 2. Rocchio Оптимальный запрос для разделения релевантных и нерелевантных документов

Этот алгоритм [7] прост в реализации, эффективен в вычислениях, быстро обучаем и имеет механизм обратной связи по релевантности, но низкую точность классификации. Линейная комбинация слишком проста для классификации, а константы α и β являются эмпирическими. Это широко используемый алгоритм обратной связи по релевантности, работающий в модели вектор-

ного пространства. Исследователи использовали вариант алгоритма Роккио в контексте машинного обучения, т. е. для изучения профиля пользователя из неструктурированного текста документа.

K-ближайший сосед (k-NN)

Алгоритм k-ближайших соседей (k-NN) [8] используется для проверки степени сходства между документами и k обучающими данными и для хранения определенного количества классификационных данных, тем самым определяя категорию тестовых документов. Этот метод представляет собой алгоритм мгновенного обучения, который классифицирует объекты на основе ближайшего пространства признаков в обучающем наборе [8]. Учебные наборы отображаются в многомерном пространстве признаков. Пространство признаков разделено на области в зависимости от категории обучающей выборки. Точка в пространстве признаков относится к определенной категории, если она является наиболее часто встречающейся категорией среди k ближайших обучающих данных. Обычно Евклидово расстояние обычно используется при вычислении расстояния между векторами. Ключевым элементом этого метода является наличие меры подобия для идентификации соседей конкретного документа [8]. Фаза обучения состоит только из сохранения векторов признаков и категорий обучающего набора. На этапе классификации вычисляются расстояния от нового вектора, представляющего входной документ, до всех сохраненных векторов, и выбираются k ближайших выборок. Аннотированная категория документа прогнозируется на основе ближайшей точки, которая была присвоена определенной категории.

$$\text{argmax}_i \sum_{j=1}^{k} sim\left(D_j|D\right) * \delta\left(C\left(D_j\right),i\right) \qquad (2)$$

Вычислите сходство между тестовым документом и каждым соседом и назначьте тестовый документ классу, который содержит большинство соседей. Рис.3.

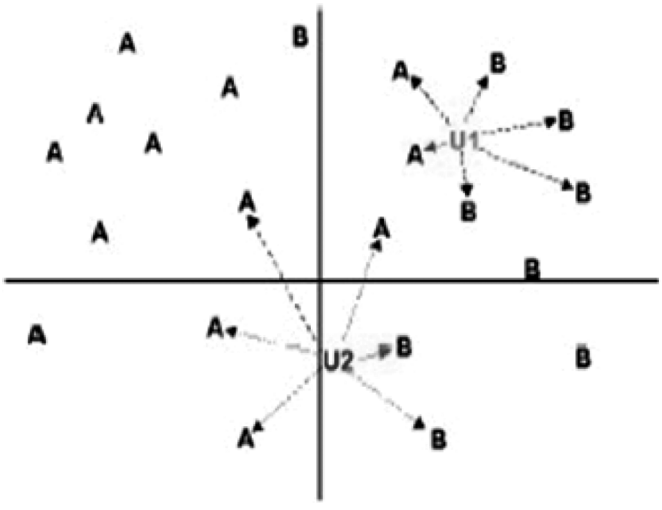

Рис. 3. k-ближайший сосед

Этот метод эффективен, непараметричен и прост в реализации. По сравнению с алгоритмом Роккио учитываются более локальные характеристики документов, однако время классификации велико и сложно найти оптимальное значение k. т. е. для анализа k-NN и алгоритма Rocchio в [6] выявлены некоторые недостатки каждого из них. В [6] предложен новый алгоритм, который включает взаимосвязь тезаурусов, основанных на понятиях, с категоризацией документов с использованием классификатора k-NN, в то время как [10] представляет использование фраз в качестве основных признаков в задаче классификации электронной почты и выполнили обширную эмпирическую оценку с использованием больших коллекций электронной почты и протестировали с тремя алгоритмами классификации текста, а именно, наивным байесовским классификатором и двумя классификаторами k-NN, использующими взвешивание и сходство TF-IDF соответственно. Метод k-ближайших соседей отличается своей простотой и широко используется для классификации текстов. Этот метод хорошо работает даже при решении задач классификации с многокатегоризированными документами. Основным недостатком этого метода является то, что он использует все функции расчета расстояния и делает метод интенсивным с точки зрения вычислений, особенно при увеличении размера обучающей выборки. Кроме того, точность классификации k-ближайших соседей сильно снижается из-за наличия зашумленных или нерелевантных признаков.

Дерево решений

Дерево решений перестраивает ручную категоризацию учебных документов, создавая четко определенные истинные/ложные запросы в форме древовидной структуры. В структуре дерева решений листья представляют соответствующую категорию документов, а ветви представляют соединения признаков, которые ведут к этим категориям. Хорошо организованное дерево решений может легко классифицировать документ, поместив его в корневой узел дерева и позволив ему пройти через структуру запроса, пока он не достигнет определенного листа, который представляет собой цель классификации документа.

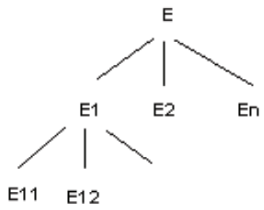

Рис. 4. Дерево решений

Метод классификации дерева решений отличается от других инструментов поддержки принятия решений рядом преимуществ. Основным преимуществом дерева решений является его простота в понимании и интер-

претации даже для неопытных пользователей. Кроме того, объяснение данного результата может быть легко воспроизведено с использованием простых математических алгоритмов и обеспечивает консолидированное представление логики классификации, что является полезной информацией о классификации.

Классификация правил принятия решений

Метод классификации правил принятия решений использует вывод на основе правил для классификации документов по их аннотированным категориям [10, 11]. Алгоритмы создают набор правил, описывающих профиль для каждой категории. Правила обычно строятся в формате «ЕСЛИ условие, ТО заключение», где часть условия заполняется признаками категории, а часть заключения представлена названием категории или другим правилом, подлежащим проверке. Затем набор правил для определенной категории создается путем объединения каждого отдельного правила из той же категории с логическим оператором, обычно использующим «и» и «или». Во время задач классификации не обязательно должно выполняться каждое правило в наборе правил. В случае обработки набора данных с большим количеством признаков для каждой категории рекомендуется реализация эвристики, чтобы уменьшить размер набора правил, не влияя на производительность классификации. В [9] представлен гибридный метод обработки на основе правил и нейронных сетей с обратным распространением для фильтрации спама. Вместо использования ключевых слов в этом исследовании используется поведение рассылки спама в качестве признаков для описания электронных писем.

Выводы

В данной статье представлен обзор подходов к машинному обучению и методов представления документов. Представлен анализ методов отбора признаков и алгоритмов классификации. В ходе исследования было подтверждено, что информационная прибыль и статистика хи-квадрат являются наиболее часто используемыми и хорошо работающими методами для выбора признаков, однако многие другие методы FS недавно были предложены в качестве одиночных или гибридных методов, показали хорошие результаты и нуждаются в дальнейшем изучении для повышения эффективности. процесс классификации.

Для автоматической классификации документов было предложено несколько алгоритмов или комбинации алгоритмов в качестве гибридных подходов. Среди этих алгоритмов SVM, NB, kNN и их гибридная система с комбинацией различных других алгоритмов и методов выбора признаков показаны наиболее подходящими в существующей литературе. Однако NB хорошо справляется с фильтрацией спама и категоризацией электронной почты, требует небольшого количества обучающих данных для оценки параметров, необходимых для классификации. Наивный Байес хорошо работает с числовыми и текстовыми данными, его легко реализовать по сравнению с другими алгоритмами, однако предположение об условной независимости нарушается реальными данными и работает очень плохо, когда признаки сильно коррелированы и не учитывают частоту вхождения слов.

ЛИТЕРАТУРА

1. А. Дасгупта, «Методы выбора признаков для классификации текстов». Материалы 13-й международной конференции ACMSIGKDD по открытию знаний и анализу данных, стр. 230–239, 2017.
2. Рагхаван, П., С. Амер-Яхия и Л. Гравано, ред., «Структура в тексте: извлечение и эксплуатация». В. Материалы 7-го международного семинара по сети и базам данных (WebDB), ACM SIGMOD/PODS 2004, ACM Press, Vol 67, 2019.
3. Корпорация Oracle, WWW, oracle.com, 2018 г.
4. Merrill Lynch, ноябрь 2020 г. Аналитика электронного бизнеса: подробный отчет. 2020.
5. Pegah Falinouss «Прогнозирование тренда акций с использованием новостных статей: подход к интеллектуальному анализу текста» Магистерская диссертация -2017.
6. Себастьяни, Ф., «Машинное обучение в автоматизированной категоризации текста» ACM Computing Surveys (CSUR) 34, стр. 1–47, 2022.
7. Андреас Хото «Краткий обзор анализа текста», 2015 г.
8. Шанг В., Хуанг Х., Чжу Х., Линь Ю., Цюй Ю. и Ван З., «Новый алгоритм выбора признаков для категоризации текста». Elsevier, science Direct Expert system with application-2006, 33(1), pp.1–5, 2016.
9. Монтанес, Э., Ферандес, Дж., Диас, И., Комбарро, Э.Ф. и Ранилья, Дж., «Показатели качества правил для выбора признаков при категоризации текста», 5-й международный симпозиум по интеллектуальному анализу данных, Германия-2019, Springer-Verlag 2019, Vol2810, стр. 589-598, 2019.
10. Ван Ю. и Ван Х.Дж., «Новый подход к выбору признаков в классификации текста», Труды 4-й Международной конференции по машинному обучению и кибернетике, IEEE-2015, том 6, стр. 3814–3819, 2015.
11. Лю, Х. и Мотода, «Извлечение, построение и выбор признаков: перспектива интеллектуального анализа данных». Бостон, Массачусетс (Массачусетс): Kluwer Academic Publishers.
12. Ли, Л.В., и Чен, С.М., «Новые методы категоризации текста на основе нового метода выбора признаков и новой меры сходства между документами», IEA/AEI, Франция, 2016 г.

13. Маномайсупат П. и Абмад К., «Выбор функций для категоризации текста с использованием самоорганизующейся карты», 2-я Международная конференция по нейронным сетям и мозгу, 2015 г., IEEE Press, том 3, стр. 1875–1880, 2015 г.
14. Ян Дж., Лю Н., Чжан Б., Ян С., Чен З., Ченг К., Фань В. и Ма В., «OCFS: оптимальное ортогональное центральное -id Выбор функции для категоризации текста». 28 Ежегодная международная конференция по исследовательскому и информационному поиску, ACM SIGIR, Баризаль, стр. 122–129, 2015 г.