Б

УДК 004.89

Е. И. Бурлаева, С. А. Зори
Государственное образовательное учреждение высшего профессионального образования «Донецкий национальный технический университет», г. Донецк
83001, г. Донецк, ул. Артёма, 58

СРАВНЕНИЕ НЕКОТОРЫХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ АНАЛИЗА ТЕКСТОВЫХ ДОКУМЕНТОВ

E. I. Burlaeva, S. A. Zori
State Educational Institution of Higher Education «Donetsk national technical University», Donetsk city
83001, Donetsk, Artema str., 58

COMPARISON OF SOME METHOD OF LEARNING METHODS FOR ANALYSIS OF TEXT DOCUMENTS

К. І. Бурлаєва, С. А. Зорі
Державна освітня установа вищої професійної освіти «Донецький національний технічний університет», м. Донецьк
83001, м. Донецьк, вул. Артема, 58

ПОРІВНЯННЯ ДЕЯКИХ МЕТОДІВ МАШИННОГО НАВЧАННЯ ДЛЯ АНАЛІЗУ ТЕКСТОВИХ ДОКУМЕНТІВ

В статье рассматривается механизм работы различных методов машинного обучения, которые используются для построения модели многоклассовой классификации слабоструктурированных текстовых документов. Задача модели состоит в отнесении документа к одному или нескольким классам, на основе анализа текстового содержания документа. Проанализирована эффективность работы различных композиций алгоритмов классификации на основе коллекции текстовых документов. Приведены результаты выполненных экспериментов, подтверждающие эффективность составленной модели, используемой для улучшения качества анализа текста. При тестировании на реальных данных наилучший результат показала композиция, состоящая из метода опорных векторов и дерева принятия решений.
Ключевые слова: Классификация документов, машинное обучение, метод опорных векторов (SVM), латентно-семантический анализ (LSA).

The article discusses the mechanism of operation of various machine learning methods that are used to build a model of multi-class classification of semi-structured text documents. The task of the model is to classify the document to one or several classes, based on the analysis of the textual content of the document. Analyzed the effectiveness of the various compositions of the classification algorithms based on a collection of text documents. The results of the performed experiments are presented, confirming the effectiveness of the model used to improve the quality of text analysis. When testing on real data, the best result was shown by the composition consisting of the support vector machine and the decision tree.
Keywords: Document classification, machine learning, support vector machine (SVM), latent-semantic analysis (LSA)

У статті розглядається механізм роботи різних методів машинного навчання, які використовуються для побудови моделі багатокласової класифікації слабоструктурованих текстових документів. Завдання моделі полягає у віднесенні документа до одного або декількох класів, на основі аналізу текстового змісту документа. Проаналізовано ефективність роботи різних композицій алгоритмів класифікації на основі колекції текстових документів. Наведено результати виконаних експериментів, що підтверджують ефективність складеної моделі, використовуваної для поліпшення якості аналізу тексту. При тестуванні на реальних даних найкращий результат показала композиція, що складається з методу опорних векторів і дерева прийняття рішень.
Ключові слова: Класифікація документів, машинне навчання, метод опорних векторів (SVM), латентно-семантичний аналіз (LSA).

Актуальность работы

Обмен информацией является неотъемлемой частью нашей жизни. В связи со стремительным развитием технического прогресса, появлением новых информационных технологий все чаще текстовые данные стали храниться и обрабатываться в электронном виде. В большинстве организаций значительная часть полезных знаний содержится в документальных базах данных. Как следствие, комфортная работа с такими большими объёмами информации затрудняется. Сложность обработки этих массивов увеличивается пропорционально их объему. Отсутствие возможности получать требуемую информацию по конкретной теме делает бесполезной большую часть накопленных данных, представленных текстами в электронном виде.

Такая ситуация обуславливает повышенный интерес к области *Text mining* – методам автоматического извлечения и обработки знаний из текстовых документов. Получение знаний в автоматическом режиме затрудняется слабой структурированностью текстов на естественном языке. Такие знания могут быть с лёгкостью извлечены экспертом, но с учётом огромного количества электронных документов их эффективная обработка человеком становится весьма затратной как по времени, так и по ресурсам.

Извлечение знаний имеет своей конечной целью информационную поддержку эксперта или автоматизированной системы при принятии проектных решений. В документах, созданных специалистами, могут быть описаны подходы к решению различных проблем, рекомендации к подбору параметров и прочие знания, полезные в различных областях деятельности организации.

Использование автоматической текстовой классификации позволяет сократить время на обработку электронных документов, но также не стоит забывать об особенностях языка, на котором составлены документы.

Востребованность в эффективном решении задачи автоматической классификации ЕЯ-текстов привела к бурному развитию новых методов и повышению эффективности уже существующих подходов.

Цель данной статьи – анализ эффективности применения композиций методов машинного обучения для задач классификации ЕЯ-текстов.

Далее будут рассмотрены различные композиции алгоритмов машинного обучения, используемые для классификации текстов, с точки зрения качества анализа текстовых документов, а также проблемы, которые возникают при применении этих методов.

Основное содержание работы

Процедура автоматической классификации текстов обычно включает две основные части: представление текстов в виде векторов признаков и построение классификатора на созданном массиве векторов. Необходимость представления текстов в виде векторов признаков определяется тем обстоятельством, что все методы классификации требуют, чтобы классифицируемые объекты были представлены в виде последовательностей чисел одинакового размера и одинакового формата.

Стандартное представление документа, используемое в текстовой классификации, является векторной моделью, представленной при помощи статистической меры – *tf-idf*. Она используется для оценки важности слова в контексте документа, входящего в некоторый текстовый корпус. Согласно определению, данная мера есть вес некоторого слова, который пропорционален количеству употребления этого слова в документе и обратно пропорционален частоте употребления слова в других документах коллекции [1] и вычисляется по формуле:

$$tf - idf(w, d, D) = tf(w, d) \times idf(w, D), \quad (1)$$

где *tf* (*term frequency* – частота терма) – отношение числа вхождения некоторого терма к общему количеству термов документа. Таким образом, оценивается важность терма w в пределах отдельного документа d и *D* – число документов в корпусе [1]. Частота слова оценивает важность слова w_i в пределах отдельного документа:

$$tf(w,d) = \frac{n_i}{\sum_k n_k}, \quad (2)$$

где n_i – число вхождений слова i в документ;

$\sum_k n_k$ – общее число слов в данном документе.

idf (*inverse document frequency* – обратная частота документа) – инверсия частоты, с которой некоторое слово встречается в документах коллекции. Учёт *idf* уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение *idf* [1]:

$$idf(w,D) = \log\frac{|D|}{|(d_i \supset w_i)|}, \quad (3)$$

где $|D|$ – количество документов в корпусе;

$|(d_i \supset w_i)|$ – количество документов, в которых встречается слово w_i.

Когда *tf-idf* функция применена ко всем словам во всех документах корпуса, слова можно отсортировать по полученным весам. Более высокий *tf-idf* вес говорит о том, что слово важно для данного документа, и в то же время достаточно редко употребляется в других документах корпуса. Это зачастую можно интерпретировать как знак того, что слово является важным для данного конкретного документа и может быть использовано, чтобы точно описать документ. Статистическая мера *tf-idf* предоставляет хорошую эвристику для определения кандидатов в ключевые слова, и этот метод (и многие его модификации) за годы исследований показал свою эффективность [2]. Данный метод в связи с эффективностью и простотой продолжает активно применяться для построения векторной модели, которая будет использоваться для классификации. Рассмотрим некоторые методы классификации.

Алгоритм классификации SVM, предложенный В. Н. Вапником [3], является контролируемым алгоритмом обучения, и принадлежит к группе методов детерминистского подхода. Это двоичный линейный классификатор, который отделяет позитивный и негативный примеры в тестовом наборе. Метод ищет гиперплоскости, отделяющие положительные примеры от отрицательных примеров, гарантируя, что граница между ближайшими позитивами и негативами является максимальной.

Метод SVM базируется на таком постулате [4]: наилучшая разделяющая плоскость – это та, которая максимально далеко отстоит от ближайших до нее точек обоих классов. То есть задача метода SVM состоит в том, чтобы найти разделяющую полосу с максимальной шириной, поскольку, чем шире полоса, тем увереннее можно классифицировать объекты, соответственно, в методе SVM считается, что самая широкая полоса является наилучшей. Границами полосы являются две параллельные гиперплоскости с направляющим вектором *w*. Точки, ближайшие к разделяющей гиперплоскости, расположены точно на границах полосы, при этом сама разделяющая гиперплоскость проходит ровно посередине полосы (рис. 1).

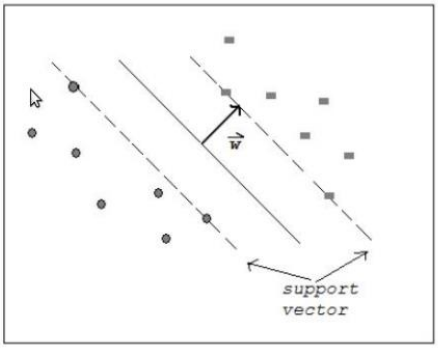

Рисунок 1 – Пример результата работы метода SVM

Еще одним методом машинного обучения является латентно-семантический анализ (*LSA, Latent Semantic Analysis*) – это теория и метод для извлечения контекстно-зависимых значений слов при помощи статистической обработки больших наборов текстовых данных [5]. LSA был запатентован в 1988 году и относится к классификаторам, основанным на функциях подобия. В области информационного поиска данный подход называют латентно семантическим индексированием (*LSI, Latent Semantic Indexing*). LSA также работает с векторным представлением типа «мешка слов» текстовых единиц. Текстовый корпус представляется в виде числовой матрицы «слово-текст», строки которой соответствуют словам, а столбцы – текстовым единицам.

Объединение слов в темы и представление текстовых единиц в пространстве тем осуществляется путем применения к данной матрице одного из матричных разложений. Наиболее популярными являются: сингулярное разложение [5] и факторизация неотрицательных матриц [6]. Согласно теореме о сингулярном разложении, любая вещественная прямоугольная матрица может быть разложена на произведение трех матриц: $A = USV^T$, где $A \in R^{n \times m}$, матрицы $U \in R^{n \times k}$ и $V \in R^{m \times k}$ – ортогональные, а $S \in R^{k \times k}$ – диагональная матрица, значения на диагонали которой называются сингулярными значениями матрицы *A*, V^T – транспонированная матрица.

Таким образом, каждое слово и текст представляются при помощи векторов в общем пространстве размерности *k* – пространстве гипотез. Сходство между любой комбинацией слов и/или текстов легко вычисляется при помощи скалярного произведения векторов. Как правило, выбор *k* зависит от поставленной задачи и подбирается эмпирически. Если выбранное значение *k* слишком велико, то метод теряет свою мощность и приближается по характеристикам к стандартным векторным методам. Слишком маленькое значение *k* не позволяет улавливать различия между похожими словами или текстами [4].

Дерево принятия решений представляет собой простой классификатор и широко используется в задачах классификации текстов.

«Деревья решений» (классификации) – это метод, позволяющий предсказывать принадлежность наблюдений или объектов к тому или иному классу категориальной зависимой переменной в соответствии со значениями одной или нескольких продикторных переменных [7].

Итак, дерево решений, подобно его «прототипу» из живой природы, состоит из «ветвей» и «листьев». Ветви (ребра графа) хранят в себе значения атрибутов, от которых зависит целевая функция; на листьях же записывается значение целевой

функции. Существуют также и другие узлы – родительские и потомки – по которым происходит их разветвление (рис. 2).

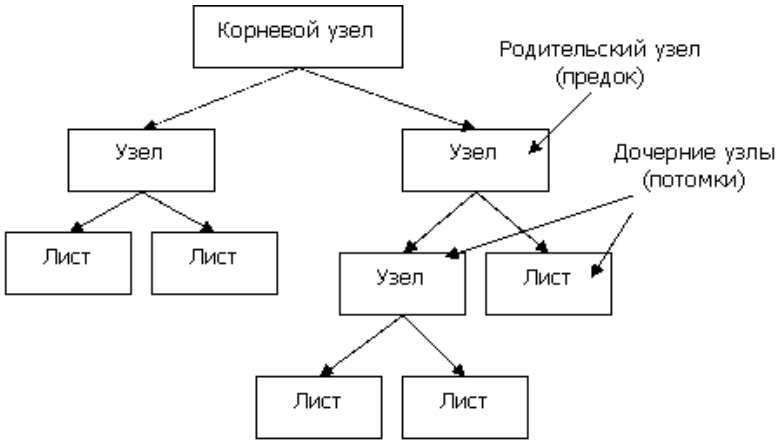

Рисунок 2 – Пример дерева решений

Один из способов автоматического построения деревьев решений заключается в последовательном разбиении множества обучающих документов на классы до тех пор, пока в классе не останется документов, определенных только в одну из категорий. На каждом этапе в качестве узла дерева выбирается терм содержащий множество всех возможных значений и определяется условие для ветвей, затем множество документов разбивается на два класса, каждый из которых имеет свои условия.

Обычно построенное дерево решений является сильно детализированным (эффект переобучения), поэтому применяются различные алгоритмы усечения дерева. Широкое применение получили алгоритмы ID3 [8] и C4.5 [8].

Оценкой качества использования методов автоматической классификации текстов является сравнение результатов их тестовых испытаний с теми оценками, которые дают этим методам эксперты. При этом «идеальным» алгоритмом считается тот, для которого выводы, сделанные по результатам тестирования, согласуются с мнением экспертов оценщиков [9].

Одним из показателей критериев при оценке качества, которое реализуется с использованием программной реализации метода классификатора, является точность. Точность (*precision*) классификации в пределах класса – это доля найденных классификатором документов, действительно принадлежащих данному классу, относительно всех документов, которые система отнесла к этому классу [10]. Точность классификации определяется следующим образом: считается отношение количества правильно классифицированных документов, к общему количеству классифицируемых документов.

Сравнение классификаторов при использовании в их составе различных методов является довольно сложной задачей по причине того, что разные входные данные могут приводить к различным результатам. Чтобы провести сравнение различных композиций, необходимо выполнить их построение и вычисление эффективности на одинаковых наборах документов для обучения и тестирования.

Для апробации предложенных методов классификации был создан тестовый корпус, состоящий из пяти тысяч текстовых документов, который описывает факты узкой предметной области. В качестве исходных данных для классификации исполь-

зуется массив текстовых документов на русском языке. Каждый документ подвергается форматированию. То есть преобразованию данных к единому формату с сохранением основного логически-структурного содержания информации. Также каждый документ подвергается преобразованию при помощи формулы *tf-idf*, в результате чего мы получаем интересующее нас описание документа – вектор признаков, который и подается на вход классификатору. Для каждого документа, в соответствии с выбранной метрикой, рассчитывается вероятность его принадлежности к каждому классу, и документ относится к соответствующей тематике.

На имеющихся исходных данных были протестированы следующие классификаторы и их композиции, которые состоят из:

– преобразование текста в вектор с помощью формулы *tf-idf*. Использование вектора для классификации при использовании метода support vector machine, которую условно будем обозначать T-I+SVM;

– преобразование текста в вектор с помощью формулы *tf-idf*. Использование вектора для классификации при использовании метода *Latent Semantic Analysis*, которую условно будем обозначать T-I+ LSA;

– преобразование текста в вектор с помощью формулы *tf-idf*. Использование вектора для классификации при использовании метода дерево принятия решений, которую условно будем обозначать T-I+D;

– преобразование текста в вектор с помощью формулы *tf-idf*. Использование вектора для классификации при помощи композиции методов *support vector machine* и дерево принятия решений, которую условно будем обозначать T-I+SVM+D;

– преобразование текста в вектор с помощью формулы *tf-idf*. Использование вектора для классификации при помощи композиции методов *Latent Semantic Analysis* и дерево принятия решений, которую условно будем обозначать T-I+LSA+D;

В зависимости от того, какая композиция будет использована, будет меняться и качество классификации, что представлено в табл. 1. В ней приведены данные проведенных экспериментов.

Таблица 1 – Точность классификации документов

Название композиций	Точность, %
T-I+SVM	88
T-I+LSA	85
T-I+D	80
T-I+SVM+D	93
T-I+LSA+D	91

В результате серии экспериментов классификаторы, тестируемые на русскоязычных текстовых документах, показали себя хорошо. Лучший результат был достигнут при использовании композиции, состоящей из метода опорных векторов и дерева принятия решений (T-I+SVM+D). Процент правильно классифицированных документов довольно высок, однако он ниже, чем при использовании этих же классификаторов на англоязычном тексте [11]. При этом наилучшим методом классификации считается тот, для которого выводы, сделанные системой, согласуются с мнением экспертов оценщиков. Стоит отметить, что использование данных методов классификации на русскоязычных текстах напрямую зависит не только от качества работы и гибкости выбранного классификатора, но и от морфологии языка, на котором составлены тексты, содержащиеся в массиве документов.

Выводы

В результате проведения серии экспериментов и основываясь на точности классификации, можно прийти к перечисленным далее выводам:

– Рассмотрена задача автоматической классификации текстовых документов.

– Точность комбинированного метода классификации оказалась выше относительно остальных методов. А именно, композиции, объединяющей в себе два классификатора – T-I+SVM+D.

Направления дальнейших исследований можно сформулировать следующим образом:

– Стоит обратить внимание на то, что подбор большего количества признаков для классификации документов при проведении экспериментов может улучшить качество классификации.

– Построение дополнительной обработки, такой как морфологический анализ, просто необходимо в связи с тем, что русский язык является флективным, что может улучшить качество работы классификатора.

Список литературы

1. Епрев А. С. Автоматическая классификация текстовых документов [Текст] / А. С. Епрев // Математические структуры и моделирование. – 2010. – № 21. – С. 65–81.
2. Недильченко О. С. Этапы и методы автоматического извлечения ключевых слов [Текст] / О. С. Недильченко // Молодой ученый. – 2017. – № 22. – С. 60.
3. Choosing Multiple Parameters for Support Vector Machines [Текст] / O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee // Machine Learning. – 2002. – Vol. 46, № 1. – P. 131–159.
4. Landauer T. K. An Introduction to Latent Semantic Analysis [Электронный ресурс] / T. K. Landauer, P. Foltz, D. Laham // Discourse Processes. – 1998. – Vol. 25. – P. 259–284.
   URL: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf (дата обращения 23.10.2018).
5. Kim H. Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method [Электронный ресурс] / H. Kim, H. Park // SIAM Journal on Matrix Analysis and Applications, 2008. – Vol. 30, №. 2. – P. 713–730.
   URL: http://www.cc.gatech.edu/~hpark/papers/simax-nmf.pdf (дата обращения 23.10.2018).
6. Воеводин В. В. Матрицы и вычисления. [Текст] / В. В. Воеводин, Ю. А. Кузнецов. – М. : Наука, 1984. – С. 320.
7. Дорогов В. Г. Введение в методы и алгоритмы принятия решений [Текст] : Учебное пособие / В. Г. Дорогов, Я. О. Теплова. – М. : ИД ФОРУМ, ИНФРА-М, 2012. – С. 240.
8. Паклин Н. Б. Бизнес-аналитика: от данных к знаниям [Текст] : Учебное пособие / Н. Б. Паклин, В. И. Орешков. – 2-е изд. – СПб : Питер, 2013. – С. 706.
9. Агеев М. С. Официальные метрики РОМИП'2004 [Текст] / М. С. Агеев, И. Е. Кураленок // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004). – Пущино, 2004. – С. 214.
10. Жизняков А. Л. Влияние изменения размерности вектор-контуров изображений на их меру близости [Текст] / А. Л. Жизняков, В. В. Зуев // Вопросы радиоэлектроники. – 2010. – Т. 1., № 1. – С. 165–170.
11. Ageev M. Text Categor ization Tasks for Large Hierarchial Systems of Categories [Текст] / M. Ageev, B. Dobrov, N. Loukachevitch // SIGIR 2002 Workshop on Operational Text Classification Systems / Eds. F. Sebastiani, S. Dumas, D. D. Lewis, T. Montgomery, I. Moulinier. – Univ. of Tampere, 2002. – P. 49–52.
12. Павлыш В. Н. Проект построения алгоритма классификации текстовых документов [Текст] / В. Н. Павлыш, Е. И. Бурлаева // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2017. – № 4 (7). – С. 24–32.
13. Павлыш В. Н. Задача классификации информации при формировании баз данных в компьютерных обучающих системах [Текст] / В. Н. Павлыш, С. А. Зори, Е. И. Бурлаева // Проблемы искусственного интеллекта. – Донецк : ГУ ИПИИ. – 2018. – № 4 (11). – С. 71–81.

References

1. Yeprev A. S. Avtomaticheskaya klassifikatsiya tekstovykh dokumentov [Automatic Classification of Text Documents]. *Matematicheskiye struktury i modelirovaniye* [Mathematical Structures and Modeling], 2010, No. 21. pp. 65–81.
2. Nedil'chenko O. S. Etapy i metody avtomaticheskogo izvlecheniya klyuchevykh slov [Stages and Methods for Automatic Keyword Extraction]. *Molodoy uchenyy* [Young Scientist], 2017, No. 22, p. 60.
3. Chapelle O., Vapnik V., Bousquet O., Mukherjee S. Choosing Multiple Parameters for Support Vector Machines. *Machine Learning*, 2002, Vol. 46, No. 1, pp. 131–159.
4. Landauer T. K., Foltz P., Laham D.An Introduction to Latent Semantic Analysis. *Discourse Processes.* 1998, Vol. 25, pp. 259–284, URL: http://lsa.colorado.edu/papers/dp1.LSAintro.pdf (дата обращения 23.10.2018).
5. Kim H., Park H. Nonnegative Matrix Factorization Based on Alternating Nonnegativity Constrained Least Squares and Active Set Method. *SIAM Journal on Matrix Analysis and Applications*, 2008. Vol. 30, No. 2. P. 713–730. URL: http://www.cc.gatech.edu/~hpark/papers/simax-nmf.pdf (дата обращения 23.10.2018).
6. Voyevodin V. V., Kuznetsov Yu. A. *Matritsy i vychisleniya* [Matrices and calculations], M., Nauka, 1984, P. 320.
7. Dorogov V. G. Teplova Ya. O. *Vvedeniye v metody i algoritmy prinyatiya resheniy* [Introduction to methods and decision-making algorithms: Tutorial]: Uchebnoye posobiye, M., ID FORUM, INFRA-M, 2012. P. 240.
8. Paklin N.B., Oreshkov V.I. *Biznes-analitika: ot dannykh k znaniyam* [Business Intelligence: From Data to Knowledge], Uchebnoye posobiye, 2-ye izd, SPb, Piter, 2013, P. 706.
9. Ageyev M.S., Kuralenok I.Ye. Ofitsial'nyye metriki ROMIP'2004 [Official metrics ROMIP'2004]. *Rossiyskiy seminar po Otsenke Metodov Informatsionnogo Poiska* (ROMIP 2004) [Russian Seminar on Evaluation of Information Retrieval Methods], Pushchino, 2004, P. 214.
10. Zhiznyakov A.L., Zuyev V.V. Vliyaniye izmeneniya razmernosti vektor-konturov izobrazheniy na ikh meru blizosti [The effect of changing the dimension of the vector contours of images on their proximity measure]. *Voprosy radioelektroniki* [Questions radioelectronics], 2010, Vol. 1, No 1, pp. 165–170.
11. Ageev M., Dobrov B., Loukachevitch N. Text Categor ization Tasks for Large Hierarchial Systems of Categories. *SIGIR 2002 Workshop on Operational Text Classification Systems* / Eds. F. Sebastiani, S. Dumas, D. D. Lewis, T. Montgomery, I. Moulinier. Univ. of Tampere, 2002. P. 49–52.
12. Pavlysh V. N., Burlayeva Ye. I. Proyekt postroyeniya algoritma klassifikatsii tekstovykh dokumentov [Draft of the algorithm for the classification of text documents]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2017, no. 4 (7), pp. 24–32.
13. Pavlysh V. N., Zori C. A., Burlayeva Ye. I. Zadacha klassifikatsii informatsii pri formirovanii baz dannykh v komp'yuternykh obuchayushchikh sistemakh [The task of classifying information when forming databases in computer-based learning systems]. *Problemy iskusstvennogo intellekta* [Problems of Artificial Intelligence], 2018, no. 4 (11), pp. 71–81.

RESUME

E. I. Burlaeva, S. A. Zori

Comparison of some machine learning methods for analyzing text documents

Introduction: The article discusses the mechanism of operation of various methods of machine learning, which are used to construct a model of multi-class classification of semi-structured text documents. The task of the model is to classify the document to one or several classes, based on the analysis of the textual content of the document. Analyzed the effectiveness of the various compositions of the classification algorithms based on a collection of text documents. The results of the performed experiments are presented, confirming the effectiveness of the model used to improve the quality of text analysis. When testing on real data, the best result was shown by the composition consisting of the support vector machine and the decision tree.

Main content: The use of automatic textual classification allows to reduce the time for processing electronic documents, but also you should not forget about the features of the language in which the documents are written.

The demand for an effective solution of the problem of automatic classification of NLText texts led to the rapid development of new methods and an increase in the effectiveness of existing approaches.

Under the text classification refers to the assignment of text information to one or more categories in accordance with certain characteristics.

The purpose of this article is to analyze the effectiveness of the use of compositions of machine learning methods for the classification of NL-texts.

A successful, high-quality algorithm is well-chosen all the "building blocks" of the algorithm.

The task of choosing the model of the classifier of documents depends on the subject area and the characteristics of the texts under consideration. Implementations of existing methods of machine learning are universal and do not take into account the specifics of the subject area.

Of interest in this paper are the quality of classification when using different compositions of machine learning algorithms, in terms of the quality of the analysis of text documents, as well as the problems that arise when applying these methods.

Results: When comparing the quality of work of different combinations of the algorithm, the composition of existing algorithms was used. The best result was obtained by the layout of methods, which showed the highest accuracy in the processing of Russian-language text documents. To construct this composition, two classifiers were used, namely, support vector machine and decision tree.

Conclusion: The key is to get the idea of combining several classifiers in the so-called composition of the algorithm to achieve higher accuracy in the classification of objects. But it is also worth noting that the selection of more features for the classification of documents during experiments can improve the quality of the classification. It is also worth considering the fact that text documents are written in Russian, the morphology of which has its own characteristics.

РЕЗЮМЕ

Е. И. Бурлаева, С. А. Зори

Сравнение некоторых методов машинного обучения для анализа текстовых документов

Введение: В статье рассматривается механизм работы различных методов машинного обучения, которые используются для построения модели многоклассовой классификации слабоструктурированных текстовых документов. Задача модели состоит в отнесении документа к одному или нескольким классам, на основе анализа текстового содержания документа. Проанализирована эффективность работы различных композиций алгоритмов классификации на основе коллекции текстовых документов. Приведены результаты выполненных экспериментов, подтверждающие эффективность составленной модели используемой для улучшения качества анализа текста. При тестировании на реальных данных наилучший результат показала композиция, состоящая из метода опорных векторов и дерева принятия решений.

Основное содержание: Использование автоматической текстовой классификации позволяет сократить время на обработку электронных документов, но также не стоит забывать об особенностях языка, на котором составлены документы.

Востребованность в эффективном решении задачи автоматической классификации ЕЯ-текстов привела к бурному развитию новых методов и повышению эффективности уже существующих подходов.

Под текстовой классификацией понимается отнесение текстовой информации к одной или нескольким категориям в соответствии с определенными признаками.

Цель данной статьи – анализ эффективности применения композиций методов машинного обучения для задач классификации ЕЯ-текстов.

Успешный, качественный алгоритм – это удачно подобранные все «кирпичики» алгоритма.

Задача выбора модели классификатора документов зависит от предметной области и характеристик рассматриваемых текстов. Реализации существующих методов машинного обучения являются универсальными и не учитывают специфики предметной области.

Интерес в данной работе представляют качество классификации при использовании разных композиции алгоритмов машинного обучения, с точки зрения качества анализа текстовых документов, а также проблемы, которые возникают при применении этих методов.

Результаты: При сравнении качества работы разных комбинаций алгоритма была использована композиция существующих алгоритмов. Наилучший результат получила компоновка методов, показавшая наивысшую точность при обработке русскоязычных текстовых документов. Для построения этой композиции были использованы два классификатора, а именно *support vector machine* и дерева принятия решений.

Заключение: Ключевой становится идея объединения нескольких классификаторов в так называемую композицию алгоритма для достижения более высокой точности классификации объектов. Но также стоит отметить, что подбор большего количества признаков для классификации документов при проведении экспериментов может улучшить качество классификации. А также стоит учесть тот факт, что текстовые документы составлены на русском языке, морфология которого имеет свои особенности.

Статья поступила в редакцию 03.09.2018.