DOI 10.37882/2223–2966.2021.01.05

СРАВНЕНИЕ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ КЛАССИФИКАЦИИ ИНТЕРНЕТ-ПУБЛИКАЦИЙ

COMPARISON OF MACHINE LEARNING ALGORITHMS OF INTERNET PUBLICATIONS CLASSIFICATION

A. Barsolevskaya
D. Kondrashkin
V. Samoylov
A. Tsaregorodtsev

Summary. The Internet is a vast source of information on Earth. You can find almost everything you want online. Information on the Internet is displayed in various formats and types, including text documents, videos and photos. Nonetheless, gathering useful information without using some web-based utilities can be a daunting task. In this scenario, web mining may come in handy. This method provides a tool that simplifies the extraction process of the required data from Internet resources. Many studies have focused on the issue of highly accurate classification of web pages and publications. This study evaluates several supervised learning algorithms to identify categories and classify social media posts. During the course of the research, following machine learning algorithms to compare the effectiveness of solving social network user publication classification problems were used: Random Forest, Neural Networks, Dimensionality Reduction, AdaBoost.

Keywords: classification of Internet user publications; social networks; classification of web pages; data analysis; random forest; neural networks; dimensionality reduction; AdaBoost.

Барзолевская Анна Федоровна
ФГБОУ ВО «Московский Государственный Лингвистический Университет», Москва
a.barzolevskaia@gmail.com
Кондрашкин Дмитрий Александрович
ФГБОУ ВО «Московский Государственный Лингвистический Университет», Москва
jakekondr@gmail.com
Самойлов Вячеслав Евгеньевич
К.т.н., ФГБОУ ВО «Московский Государственный Лингвистический Университет», Москва
v.samoilov@linguanet.ru
Царегородцев Анатолий Валерьевич
Д.т.н, профессор, ФГБОУ ВО «Московский Государственный Лингвистический Университет», Москва
avtsaregorodtsev@linguanet.ru

Аннотация. Интернет представляет собой один из крупнейших источников информации в мире. Можно сказать, что любая тема, на которую мы размышляем, может быть найдена в сети. Информация в Интернете представлена в разных формах и типах, например, в текстовых документах, картинках и видео. Однако, сбор полезной информации без помощи некоторых веб-утилит является очень непростой задачей. В этом деле помогает веб-майнинг: данный метод предоставляет инструменты, облегчающие извлечение необходимых данных из интернет-ресурсов. Многие исследования сфокусированы на проблеме высокоточной классификации веб-страниц. В данном исследовании проводится оценка некоторых алгоритмов контролируемого обучения для выявления категорий среди публикаций в социальных сетях. В процессе исследования, для сравнения эффективности решения задач классификации публикаций пользователей социальных сетей, использованы следующие алгоритмы машинного обучения: случайный лес, нейронная сеть, снижение размерности, AdaBoost.

Ключевые слова: классификация публикаций пользователей сети Интернет; социальные сети; классификация веб-страниц; анализ данных; случайный лес; нейронные сети; снижение размерности; AdaBoost.

Введение

В сети, данные представляют собой очень важную область. Их количество постоянно растет, поэтому актуальной является задача поиска полезной информации из огромного массива данных. Общий процесс анализа данных для поиска полезной информации называется интеллектуальным анализом данных. В последние годы, значительная часть корпоративных данных хранилась в реляционных базах данных [1]. Они структурированы и легкодоступны для исследований с помощью методов интеллектуального анализа данных. Однако, характер данных поменялся с появлением Интернета. В сети содержатся различные

типы и форматы данных, такие как таблицы, XML документы, неструктурированные данные, мультимедиа, текст на веб-страницах. Такое разнообразие форматов данных помогает пользователям делиться своими идеями на веб-ресурсах, но также создаёт сложности для анализа контента.

Социальные сети — это посредник в общении между людьми. Они позволяют пользователям быстро и удобно обмениваться информацией. Однако, публикации большинства пользователей тяжело классифицировать, так как зачастую необходимо использовать семантические свойства языка, на котором написано сообщение. Это приводит к появлению потребности разработке новых методов и алгоритмов интеллектуального анализа данных.

Анализ иностранных работ, посвящённых текстовому анализу

В методах текстового анализа используются концепции из многих областей, таких как фильтрация и поиск информации, искусственный интеллект, интеллектуальный анализ текста, методы машинного обучения и так далее. В модели машинного обучения, классификатор проходит обучение на уже классифицированных примерах, изучая правила классификации. Затем этот же классификатор используется для классификации новых страниц. Примеры работ:

1. Anagnostopoulos [2] предложил систему для идентификации и категоризации веб-страниц на основе фильтрации информации. Система представляет собой трехуровневую вероятностную сеть, имеющую смещения и радиальные базисные нейроны в среднем слое, и конкурирующие нейроны в выходном слое. Таким образом, идентифицируются веб-страницы для покупки-продажи онлайн, для классификации их по соответствующим типам на основе структуры, позволяющей описывать коммерческие транзакции в сети.
2. В том же направлении работал и Feng Shen [3], предложивший новую модель классификации текста для решения проблем категоризации китайского текста веб-страниц, основанную на глубоком обучении. Сети глубокого обучения обладают отличной способностью к обучению признакам. Данная система позволяет формировать укрупнённые более подходящие для классификации объекты из комбинации низкоуровневых объектов.
3. J. Jagani [4] и M.S. Othman [5] исследовали классификацию веб-документов с использованием метода извлечения информации и машинного обучения. Были определены шесть признаков веб-документа: текст, мета-тег, заголовок (A), заголовок и текст (B), заголовок (C), мета-тег, заголовок (D), мета тег (D), текст (F). Для классификации документов был использован метод опорных векторов, в то время как радиальная базисная функция, и линейные, полиномиальные и сигмовидные ядра были применены для проверки точности классификации.
4. E. Sarac [6] представил результаты исследования доказывающие, что увеличение количества информации вызвало необходимость в точной автоматической классификации веб-страниц, в целях поддержки веб-каталогов и повышения производительности поисковых систем. Каждый тег и каждый термин на каждой веб странице можно рассматривать как признак. Цель заключалась в том, чтобы применить новейший метод оптимизации, а именно «алгоритм светлячков». Данный метод был реализован для выбора подмножеств признаков и оценки соответствия выбранных признаков.
5. A. Herrouz [7] использовал методы алгоритма Apriori и реализацию наивного байесовского классификатора. Алгоритм Apriori находит корреляцию между большими наборами элементов данных, а наивный байесовский классификатор рассчитывает вероятности использования ключевых слов среди больших датасетов.

Алгоритмы машинного обучения

Поскольку текстовый анализ осуществляется на базе машинного обучения, то необходимо рассмотреть machine learning (ML) алгоритмы.

1. **Алгоритм «Случайный лес»**

Случайный лес [8] — это алгоритм машинного обучения, заключающийся в использовании ансамбля деревьев решений. Случайный лес осуществляет классификацию не на основе одного дерева решений, а на основе комплексного прогноза, построенного деревьями решений.

Обучение алгоритма «Случайный лес» использует метод бэггинг Бреймана. Классификатор обучается на тестовом наборе данных, из которого случайным образом формируются тренировочные подвыборки с повторениями. На основании каждой подвыборки строится дерево рещений. Эта процедура приводит к повышению точности модели, поскольку снижает разброс оценки классификации.

Классификация входных данных проводится путём голосования: каждое дерево комитета относит класси-

фицируемый объект к одному из классов, и побеждает класс, за который проголосовало наибольшее число деревьев алгоритма. Они определяют класс, а затем получают среднюю оценку классификации по всем деревьям.

2. Искусственные нейронные сети (ИНС)

Взаимосвязанный набор искусственных нейронов, созданный программами, подобными работе биологического нейрона или электронных структур (электронных чипов, предназначенных для моделирования работы нейронов), использующих математическую модель для обработки информации на основе коммуникативного метода вычислений. Нейронные сети обычно состоят из простых обрабатывающих элементов, которые выполняют простую работу, но общее поведение сети определяется связями между этими различными элементами, называемыми нейронами, и индикаторами этих элементов [9]. Первое предположение об идее нейронных сетей связано с механическим действием нейронов мозга, которые можно сравнить с электрическими, биологическими сетями, которые обрабатывают информацию, содержащуюся в мозге.

Искусственные нейронные сети состоят из узлов, называемых нейронами или обрабатывающими элементами, которые соединены вместе, образуя сеть. Каждый контакт между этими узлами имеет набор значений, называемых весовыми коэффициентами, которые способствуют определению значений, получаемых в результате каждого элемента обработки, на основе входных значений этого элемента. Нейронная сеть состоит из слоев искусственных ячеек [10]: входной слой, выходной слой и слои между ними, называемые скрытыми слоями. Каждая ячейка в одном из этих слоев соединена со всеми нейронами в следующем и предшествующем слоях.

Нейроны умножают каждое входное значение от нейронов предыдущего слоя на весовой коэффициент связи этих нейронов, а затем перемножают результаты. Преобразование отличается в зависимости от типа нейрона, состояние преобразования учитывает состояние нейрона, которое передается в нейроны следующего слоя.

3. Снижение размерности

Два компонента снижения размерности — отбор и проекция признаков. В первом происходит поиск подмножества исходного набора переменных или признаков, чтобы получить меньшее подмножество, которое можно использовать для моделирования проблемы. Во втором данные в пространстве большой размерности сокращаются до пространства меньшей размерности.

Снижение размерности может быть линейным или нелинейным, в зависимости от используемого метода. Этот алгоритм особенно эффективен при большом количестве признаков у объекта классификации.

4. Алгоритм AdaBoost

AdaBoost, сокращенно «Adaptive Boosting», представляет собой «мета-алгоритм машинного обучения», это тип «ансамблевого обучения», при котором различные ученики используются для создания более сильного алгоритма обучения. AdaBoost — один из самых эффективных алгоритмов контролируемого обучения за последние несколько лет [11]. В большинстве случаев он используется с несколькими альтернативными алгоритмами обучения («слабые ученики») для повышения производительности. AdaBoost работает, выбирая базовый алгоритм и итеративно улучшая его, учитывая неправильно классифицированные примеры в обучающей выборке.

Сбор и предварительная обработка данных

Сбор данных — это процесс, используемый для сканирования веб-страницы. Чтобы добиться правильного представления содержания, должна применяться предварительная обработка текста. Предварительная обработка содержит три шага (лексический анализ, токенизация строк, удаление стоп-слов и выделение корней).

Инструменты лексического анализа используются для уменьшения объема данных путем удаления любого бесполезного слова и может получать потоки полезных слов, которые можно использовать на следующих этапах предварительной обработки. С помощью токенизации каждое слово представляется в виде токена.

После того, как процесс токенизатора строки применен к стоп-словам, выполняется исключение, что означает, что местоимения, предлоги и союзы удаляются из документа, потому как не имеют никакого значения или указаний о содержании поста. Наконец, последняя часть предварительной обработки — это выделение корней — метод, используемый для минимизации слов до их морфологических частей. Процесс производит удаление суффиксов и приставок. Таким образом, можно сократить количество терминов в документе и понизить сложность классификации.

После предварительной обработки текста создаётся база данных, которая содержит все уникальные слова. Она представляет собой перечень отличительных слов, появлявшихся несколько раз. Каждое слово представляет одно признаковое описание объекта. Этот вектор

Таблица 1. Сравнительные результаты трех классификаторов

Алгоритм	Точность	Полнота	F Меры
Случайный лес	91,93%	81,17%	88,01%
Нейронная сеть	90,11%	78,35%	83,42%
Снижение размерности	87,52%	82,40%	83,40%
AdaBoost	81,64%	88,64%	82,42%

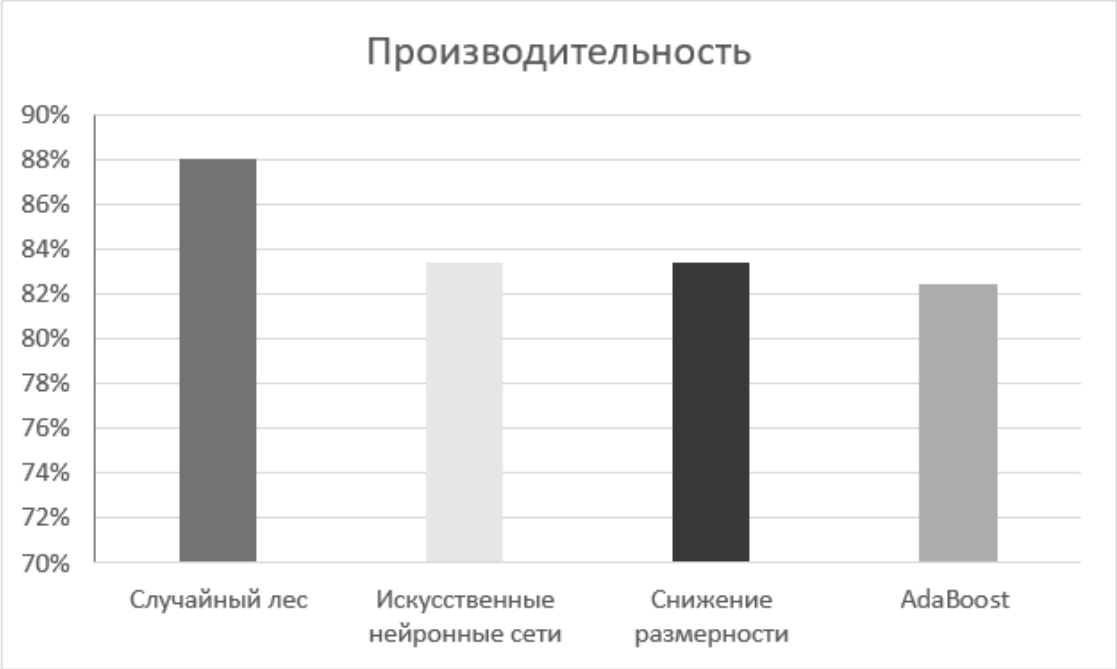

Рис. 1. Производительность алгоритмов

признаков содержит вес слов текста. Вес слова рассчитывается на основании схемы взвешивания элементов энтропии.

Схема взвешивания элементов энтропии

Энтропийный метод основан на вероятностном анализе, определяющем наиболее вероятные слова каждой категории. Он вычисляет веса слов с помощью схемы взвешивания. Весовой коэффициент рассчитывается по двум переменным: локальный весовой коэффициент L_{jk} и глобальный G_k. Схема взвешивания каждого элемента энтропии рассчитывается как $L_{jk} \times G_k$ и вычисляется по двум формулам [12]

$$\begin{cases} L_{jk} = 1 + \log\left(TF_{jk}\right), TF_{jk} > 0 \\ L_{jk} = 0, TF_{jk} = 0 \end{cases},$$

$$G_k = \frac{1+\sum_{k=1}^{n} \frac{TF_{jk}}{F_k} \log\left(\frac{TF_{jk}}{F_k}\right)}{\log n},$$

где n — количество документов в базе данных;

TF_{jk} — это частота термина для каждого слова в документе j;

F_k — это частота термина k во всём тексте.

Экспериментальные результаты

Экспериментальные данные реализации перечисленных алгоритмов приведены, чтобы показать их эффективность и различия. Используемый тип данных — публикации, которые случайным образом загружаются из социальных сетей Facebook, VKontakte и т.д. разделенные на восемь категорий, а именно: новости, события, фотографии, фотографии, видеозаписи, инфографики, прямые эфиры, опросы, этот набор данных был разде-

лен на 70% для фазы обучения и 30% для тестирования. Оценивалась эффективность классификации с использованием стандартных мер поиска информации (F-мера), он учитывает, как точность, так и полноту, как показано ниже:

$$accuracy = \frac{TP}{TP+FP},$$

где TP — количество истинно положительных результатов;

FP — количество ложных положительных результатов;

$$completeness = \frac{TP}{TP+FN},$$

где FN — количество ложноотрицательных результатов;

$$F - \text{мера} = \frac{2TP}{2TP+FP+FN}.$$

Эксперимент проводился с использованием трех алгоритмов, проверялась точность и эффективность каждого из них. Результаты представлены в таблице 1 и на рисунке 1.

Приведенные выше результаты показывают, что алгоритм «Случайный лес» имеет более высокую точность, чем нейронная сеть и AdaBoost.

Алгоритм «Случайный лес» работает с помощью обучения множества деревьев решений, каждое из которых основано на разной повторной выборке исходных обучающих данных. Создавая множество таких деревьев, а затем, усредняя их, можно значительно уменьшить разброс значений классификации по сравнению с разбросом одного дерева. На практике единственным ограничением размера «леса» является время вычисления, поскольку бесконечное количество деревьев можно обучить без постоянно увеличивающейся систематической ошибки и с постоянным уменьшением дисперсии значений.

«Слабые ученики» AdaBoost имеют высокую систематическую ошибку и низкую дисперсию.

Хотя результат алгоритма «Случайный лес» практически схож с нейронной сетью, оба алгоритма имеют сильные и слабые стороны. Здесь стоит сосредоточиться на положительных сторонах алгоритма «Случайный лес» по сравнению с нейронной сетью. Он быстрее, обычно заканчивается в течение нескольких минут, и его легче обучить, но имеет меньше параметров для настройки. Напротив, нейронная сеть имеет огромное количество параметров: количество слоев, количество нейронов в каждом слое, активация признаков, скорость обучения и т.д.

Заключение

На этапе тестирования точность классификатора алгоритма «Случайный лес» составила 88.01%, нейронной сети — 83.42%, снижения размерности — 83.4%, а AdaBoost — 82.42%. Согласно этим результатам можно сделать вывод, что алгоритм «Случайный лес» может классифицировать точнее, чем классификаторы нейронных сетей и AdaBoost. Однако, у ИНС лучшая производительность по сравнению с AdaBoost.

Для большинства архитектур нейронных сетей необходимо, чтобы данные были очень хорошо обобщены, и число входных текстов было велико. В то же время, алгоритм «Случайный лес» может обеспечить должную точность с небольшим количеством текстов.

ЛИТЕРАТУРА

1. T. Bourgeois, "Information Systems for Business and Beyond", Edition, Textbook Equity, Saylor Academy, 2014.
2. I. Anagnostopoulos, C. Anagnostopoulos, V. Loumos and E. Kayafas, "Classifying Web pages employing a probabilistic neural network", IEE Proceedings-Software, Vol.151, No.3, June 2004, PP.139–150.
3. F. Shen, X. Luo and Yi. Chen, "Text Classification Dimension Reduction Algorithm for Chinese Web Page Based on Deep Learning", International Conference on Cyberspace Technology (CCT 2013), pp. 451–456, Beijing, China, 23 Nov. 2013.
4. M.S. Othman, L.M Yusuf and J. Salim, "Web classification using extraction and machine learning techniques", In Information Technology (ITSim), 2010 International Symposium in Vol 2, PP. 765–770, Kuala Lumpur, 15–17 June 2010. (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 11, 2017 209 | P a g e www.ijacsa.thesai.org
5. E. Sarac and S. A. Ozel, "Web page classification using firefly optimization", In Innovations in Intelligent Systems and Applications (INISTA), 2013 IEEE International Symposium on, PP.1–5, Albena, Bulgarian, 19–21 June 2013.
6. M. Klassen, "A framework for search forms classification" In Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, PP.1029–1034 Seoul, Korea, 14–17 Oct. 2012.
7. K.J. Patel and K. J. Sarvakar, "Web Page Classification Using Data Mining", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No.7, July 2013, pp. 2513–2520.

8. T.K Ho, "Random decision forests." In Document Analysis and Recognition, 1995. Proceedings of the Third International Conference on, vol. 1, pp. 278–282. IEEE, 1995.
9. J. Heaton, " Artificial Intelligence for Humans: Deep learning and neural networks", Vol 3, Heaton Research, Incorporated, 323 pages, 2015.
10. N. Gupta," Artificial Neural Network", Network and Complex Systems, Vol 3, no. 1, pp. 24–28, 2013.
11. Y. Freund and R. E. Schapire."A decision-theoretic generalization of online learning and an application to boosting." In a European conference on computational learning theory, pp. 23–37. Springer, Berlin, Heidelberg, 1995.
12. Z.S. Lee, M. A. Maarof, A. Selamat and S. M. Shamsuddin, " Enhance Term Weighting Algorithm as Feature Selection Technique for Illicit Web Content Classification ", The 8th International Conference on Intelligent Systems Design and Applications, pp.145–150, Kaohsiung, Taiwan, IEEE, 26–28 Nov. 2008.

Московский государственный лингвистический университет