УДК 004

Г.В. Ломонос, асп.; Д.М. Романенко, доц.
(БГТУ, г. Минск)

КОМБИНИРОВАНИЕ МЕТОДОВ АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВОЙ ИНФОРМАЦИИ

В последние десятилетия благодаря повсеместной цифровизации и компьютеризации всех сфер жизни общества значительно возросло количество текстов доступных для анализа. Многие статьи, новостные ленты и документы теперь доступны в электронном виде. На сегодняшний день уже невозможно представить себе современную малую или среднюю организацию без электронного документооборота. Все поступающие и исходящие документы находятся в электронном виде. Чтобы успешно конкурировать с крупными организациями, малым и средним организациям следует активно внедрять цифровые технологии, сервисы и продукты, позволяющие автоматизировать повседневный ручной труд сотрудников компании, добиваясь максимальной отдачи. Это может сыграть важную роль в успехе небольшой фирмы, так как крупные организации довольно часто не уделяют должного внимания данной проблеме, в связи с тем, что имеют достаточные человеческие ресурсы и могут позволить себе до определенного момента не придавать этому значения. Внедрение системы, позволяющей автоматически классифицировать поступающие текстовые документы, позволило бы сократить расходы и повысить качество работы компании.

На сегодняшний день автоматическая классификация текста, т.е. определение принадлежности текста к некой категории в условиях постоянно возрастающего объема используемой информации является актуальной и крайне интересной задачей. Категоризация текстовых данных или классификация документов является одной из основных задач в области анализа текстовых данных и поиска информации. На текущий момент, исследовано не мало эффективных методов и алгоритмов, применяемых для классификации текстовых данных, однако часто встречается что многие из них имеют ограниченную область применения. Последние исследования в области классификации текстов показывают, что использование комбинации из некоторых стандартных методов классификации показывают намного лучшие результаты.

Важным шагом в классификации текста является классификация текстовых документов среди некоторого известного набора классов (категорий). Важно классифицировать документ придерживаясь указанной сферы, потому что проблемы возникающие при интеллекту-

альном анализе текста, такие как суммирование текста, извлечение информации, обнаружение семантических отношений и т.д., могут быть корректно решены только исходя из специфики конкретной области документа.

Задача интеллектуального анализа текста может быть выполнена с помощью двух процессов: классификации и кластеризации. В то время как кластеризация – это неконтролируемый подход к обучению, который направлен на группирование набора связанных объектов данных в кластеры на основе некоторой меры сходства или расстояния между ними. Классификация – это контролируемая форма машинного обучения, целью которой является определение категории из набора категорий, к которым относится выбранный текст [1]. Это делается на основе заранее определенного набора данных для обучения.

На текущий момент известно достаточное количество методов классификации текста отлично зарекомендовавших себя при выполнении своих задач, но в целом ни один из методов не может быть оценен как лучший или худший для классификации текста. Вероятно, это связано с тем, что каждый метод имеет свое конкретное приложение или сферу, или он ограничен определенным видом данных для классификации текста.

В последнее время ведутся работы в направлении комбинирования известных методов классификации текста, так называемый гибридный подход. Этот подход фокусируется на индивидуальных преимуществах каждого метода, отбрасывая их недостатки и проверяя работали они с пользой или нет. Таким образом, объединение методов позволяет использовать преимущества каждого из них, которые в полной мере соответствуют потребностям выполняемой классификации, сводя к минимуму индивидуальные ограничения каждого метода [2].

Методы, применяемые при классификации текста, могут быть разделены на линейные (метод опорных векторов, логистическая регрессия) и вероятностные методы (метод Байеса, метод максимальной энтропии). Впоследствии, в ходе комбинации методов, методы, принадлежащие одной и той же парадигме, могут быть объединены.

Ларки и Крофт в [3] предлагают комбинацию трех классификаторов: KNN (К Ближайших соседей), обратной связи по релевантности и байесовских классификаторов, которые будут использоваться в медицинской области для автоматического назначения кодов ICD9. Задание было выполнено сначала с каждым из классификаторов в отдельности, а затем с объединением, чтобы проверить эффективность обоих подходов. В результате, благодаря объединению методов удалось добиться намного лучших результатов. Производительность методов из-

мерялась на основе ранга документов. Это один из примеров, где классификаторы используются для ранжирования документов. Подход заключается в использовании взвешенной линейной комбинации.

Ученый доктор Беннет совместно с коллегами [4] предложил вероятностный метод объединения классификаторов таким образом, чтобы вклад классификатора зависел от его достоверности. Достоверность измеряется с помощью показателей надежности, которые связаны с областями, где классификатор может работать довольно хорошо или плохо. Вместо ранга документа показатели основаны на эффективности самого классификатора, что делает предложение более обобщенным.

Еще один из вариантов классификации основан на теории Дампстера-Шафера. Его основная цель - объединение подклассификаторов, поскольку их применение направлено на классификацию по нескольким меткам.

Ученый Дино Иса совместно с коллегами в своих двух последовательных работах [5] предложили идею о том, как мета-результаты наивной байесовской техники могут использоваться с методом опорных векторов и самоорганизующимися картами (SOM) соответственно. Формула Байеса используется для преобразования текстового документа в векторное пространство, где значения обозначают вероятности документов для любого класса в зависимости от содержащихся в нем признаков. Это называется фазой векторизации классификатора. Это общее для обоих классификаторов. Метод опорных векторов затем применяется к этой модели векторного пространства для окончательного результата классификации. Предложение улучшило точность классификации по сравнению с чисто наивным байесовским классификационным подходом. За распределениями вероятностей, полученными с помощью метода Байеса, следует этап индексации, выполняемый с помощью самоорганизующихся карт для получения случаев наилучшего соответствия. Метод опорных векторов похож на кластеризацию документов на основе меры сходства между документами, как евклидово расстояние.

В последних исследованиях ученого Фрагоса [6] также делается вывод в пользу объединения разных подходов к классификации текста. Методы, которые объединил автор, относятся к одной и той же парадигме - вероятностей. Наивные байесовские и максимальные энтропийные классификаторы использовались для тестирования в приложениях, в которых эти методы могли бы показать максимальную производительность. Для получения результатов использовались операторы слияния.

На текущий момент объединение методов классификации текста становится все более многообещающей областью исследований. Объединение методов классификации дает намного лучшие результаты, чем использование тех же методов в отдельности. Получаемые результаты исследований все более стимулируют исследования в области классификации текста с помощью комбинаций методов. Таким образом, применение в организации автоматизированных систем классификации текстовой информации, позволит сократить расходы на человеческие ресурсы и повысить производительность труда и уровень удовлетворенности клиентов.

ЛИТЕРАТУРА

1. Shai Shalev-Shwartz, Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms/ Shai Shalev-Shwartz, Shai Ben-David // Cambridge, 2014 г. / С. 449.

2. W-C. Lin, S-W. Ke, C-F. Tsai, “An intrusion detection system based on combining cluster centers and nearest neighbors”, Knowledge-Based Systems, vol. 78, 2015, С. 97.

3. M. Sharma, K. Das, M. Bilgic, B. Matthews, D. Nielsen, N. Oza, “Active Learning with Rationales for Identifying Operationally Significant Anomalies in Aviation”, Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, vol 9853, 2016, С. 359.

4. M. Panda, A. Abraham, M. Patra, “Hybrid intelligent systems for detecting network intrusions”.Security Comm. Networks,vol. 8, 2012, С. 2995.

5. Dino Isa, V. P Kallimani and Lam Hong lee, “Using Self Organizing Map for Clustering of Text Documents”, Expert System with Applications, vol. 36, no. 5, July, 2017, С. 596.

6. S. Ramasundaram, “NGramsSA Algorithm for Text Categorization”, International Journal of Information Technology & Computer Science (IJITCS), Volume 13, Issue No : 1, 2014, С. 49.