УДК 004.85: 8

Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «случайный лес»

И.С. Веретенников, Е.А. Карташев, А.Л. Царегородцев

Югорский научно-исследовательский институт информационных технологий (Ханты-Мансийск, Россия)

Evaluation of Text Materials Classification Quality Using «Random Forests» Machine Learning Algorithm

I.S. Veretennikov, E.A. Kartashev, A.L. Tsaregorodtsev

Ugra Research Institute of Information Technologies (Khanty-Mansiysk, Russia)

Представлены результаты оценки качества классификации текстовых материалов алгоритмом машинного обучения «случайный лес», реализованным в библиотеке scikit-learn. Приведено описание применяемых функций из данной библиотеки, а также параметров, которые влияют на качества классификации. Описаны основные этапы работ классификации текстовых материалов: формирование наборов материалов для обучения и контроля (обеспечение репрезентативности выборки, обработка текста, определение групп для обучения и контроля); обучение модели классификатора; тестирование модели классификатора; оценка качества полученных результатов. Осуществлена оценка качества с использованием таких характеристик, как точность (precision), полнота (recall) и F-меры работы классификатора для различных вариантов подготовки данных: сбалансированная и несбалансированная обучающие группы материалов, при этом для последней был предусмотрен вариант с преобразованием текста в набор токенов. По результатам работы определены основные направления для повышения качества классификации текстовых материалов алгоритмом машинного обучения «случайный лес».

Ключевые слова: библиотека scikit-learn, машинное обучение, классификация текстовых документов, алгоритм «случайный лес», дерево принятия решений.

The results of quality evaluation of text materials classification by the "random forests" machine learning algorithm implemented in the “scikit-learn” library are presented. Functions used in the “scikit-learn” library, as well as the parameters that affect classification quality, are described. The main stages of text materials classification are shown in the paper: the formation of sets of materials for training and control (ensuring sample representativeness, text processing, definition of groups for training and control); classifier model training; classifier model testing; quality evaluation of the obtained results. The quality evaluation is carried out using characteristics, such as precession, recall and F-measures of the classifier for various data preparation options: balanced and unbalanced training groups of materials, while the latter case is designed to convert the text into a set of tokens. Based on the results of the work, the main directions for improving quality of text materials classification by the "random forests" machine learning algorithm have been determined.

Key words: scikit-learn library, machine learning, classification of text documents, random forests classifier, decision trees.

DOI 10.14258/izvasu(2017)4-13

Введение. Увеличение объема обрабатываемой информации приводит к необходимости применения различных алгоритмов машинного обучения для ее классификации и кластеризации. Одной из информационных систем, реализующей указанный функционал, является автоматизированная информационная система поиска и анализа информации в сети Интернет (далее АИС Поиск), разработанная в Югорском научно-исследовательском институте информационных технологий [1]. В АИС Поиск на протяжении нескольких последних лет формируется набор материалов по нескольким направлениям («наркоторговля», «экстремизм», «терроризм» и т.д.), содержащих неструктурированную информацию, которая имеет следующие особенности:

— имеет разметку для отображения информации в браузере (теги), подпрограммы (скрипты) для обеспечения дополнительного функционала и т.д.;

— значимый текст является небольшим фрагментом, который не связан с остальной частью материала, содержит жаргонизмы, сокращения и орфографические ошибки;

— значимая информация может быть представлена в виде ссылки на медиаконтент (изображение, аудио- или видеозапись);

— содержит нерелевантную информацию (сведения об ошибках, отсутствии страницы или ограничении доступа, окно авторизации или приветствия и т.д.).

Указанные особенности необходимо учитывать в ходе классификации набора материалов АИС Поиск.

Задача классификации текстов [2, 3] заключается в определении логического значения для каждой пары $(d_j, c_i) \in D \times C$, где $D = \{d_1,...d_{|D|}\}$ — множество материалов, $C = \{c_1,...c_{|C|}\}$ — множество предопределенных классов. Указанное значение равно 1 (true) для пары (d_j, c_i), если материал d_j определен к классу c_i, в противном случае значение равно 0 (false).

$$\Phi(d_j, c_i) = \begin{cases} 0, & \text{если } d_j \notin c_i \\ 1, & \text{если } d_j \in c_i \end{cases} \quad (1)$$

Формализуется данная задача как аппроксимация неизвестной целевой функции $\Phi: D \times C \rightarrow \{0, 1\}$, которая определяет, как именно должны быть классифицированы материалы с помощью максимально близкой к ней функции $\Phi': D \times C \rightarrow \{0, 1\}$, которая называется классификатором.

Задача классификации текстов, решаемая АИС Поиск, предусматривает организацию бинарных классификаторов по каждому из направлений поиска материалов, которые определяются функцией $\Phi': D \rightarrow \{0, 1\}$, являющейся аппроксимацией функции $\Phi: D \rightarrow \{0, 1\}$. Данные классификаторы могут рассматриваться как множество классификаторов, которые требуется найти.

Для решения задачи АИС Поиск по автоматической классификации материалов было проведено изучение нескольких наиболее распространенных алгоритмов машинного обучения: «случайный лес» [2, 4], опорных векторов [2, 5, 6], Байеса [2, 6, 7], k-средних [8], нейронные сети [9, 10].

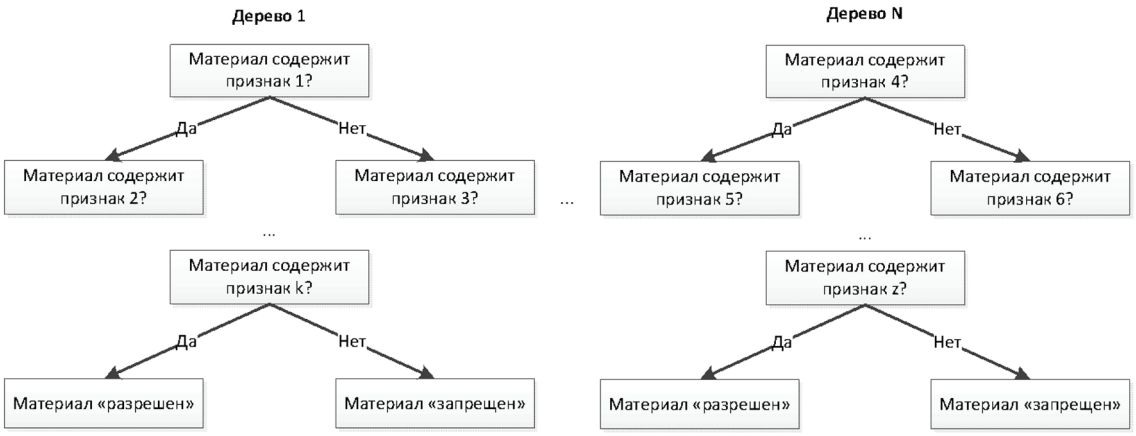
В рамках данной работы осуществлена оценка качества классификации материалов по направлениям в АИС Поиск с применением алгоритма машинного обучения «случайный лес», который также успешно применялся при решении подобных задач в других информационных системах [11]. Данный метод предусматривает:

1) построение бинарных решающих деревьев на основе обучающих данных (рис.) в соответствии с алгоритмом CART (Classification and Regression Tree), при котором каждый узел дерева при разбиении имеет только двух потомков;

2) определение класса материала по каждому из решающих деревьев;

3) выбор наиболее часто встречающегося класса.

Выполнение работ осуществлялось с использованием библиотеки scikit-learn [12], которая находится в свободном доступе, включая исходные коды, и может использоваться в других проектах на основании лицензии BSD (англ. BSD license, Berkeley Software Distribution license — программная лицензия Университета Беркли). Данная библиотека ориентирована на моделирование данных и предоставляет реализацию целого ряда методов и алгоритмов машинного обучения, в том числе наивным байесовским классификатором, нейронные сети, метод опорных векторов и алгоритм «случайный лес». На сегодняшний день команда разработчиков включает около 40 активных участников, которые на регулярной основе с 2010 г. (текущая версия 0.18.1 от 11.11.2016) при финансовой поддержке от INRIA, PARIS-SACLAY Center for Data Science, Moore-Sloan Data Science Environment, Columbia University.



Графическая схема модели «случайный лес»

Цель данной работы — оценка качества классификации текстовых материалов алгоритмом машинного обучения «случайный лес» в автоматизированной информационной системе поиска и анализа информации в интернете.

С учетом этого планируется выполнить:

Формирование наборов материалов для обучения и контроля

Из имеющихся и размеченных специалистами материалов по различным направлениям были отобраны материалы, которые содержат значимую информацию. В последующем из исходных материалов были удалены теги, скрипты, слова, которые не несут значимую для анализа информацию (предлоги, союзы и частицы), и проведена лемматизация текста (приведение слова к словарной форме). Данные работы проводились с учетом информации, представленной в работе [13]. Все полученные материалы по каждому направлению были разделены на две группы, содержащие материалы с различными отметкам. Результаты представлены в таблице 1.

Также были подготовлены материалы, над которыми лемматизация не проводилась, т.е. оставлены токены. Количественная характеристика также представлена в таблице 1.

Таблица 1

Несбалансированное количество материалов в группах по направлениям

Направление	Обучение		Контроль	
	«разрешен»	«запрещен»	«разрешен»	«запрещен»
Наркоторговля	7072	430	1750	25
Экстремизм	1083	282	863	88
Терроризм	2005	133	1230	33

В описании алгоритма машинного обучения «случайный лес» и практическом опыте его использования для классификации материалов указывается на необходимость обеспечить равное количество материалов в каждом классе на этапе обучения модели классификатора, так как это существенным образом влияет на качество результата его работы. С учетом этого путем уменьшения количества материалов с оценкой «разрешен» были сформированы группы «обучение» по каждому направлению, которые представлены в таблице 2.

Таблица 2

Сбалансированное количество материалов в группах по направлениям

Направление	Обучение		Контроль	
	«разрешен»	«запрещен»	«разрешен»	«запрещен»
Наркоторговля	220	220	8617	220
Экстремизм	168	168	1812	168
Терроризм	75	75	3176	75

Обучение модели классификатора

С учетом того, что материалы по каждому направлению будут классифицироваться независимо друг от друга, а также используется бинарная классификация («разрешен»/«запрещен»), то для группы «обучение» каждого направления с использованием функций HashingVectorizer и TfidfTransformer библиотеки scikit-learn был сформирован нормализованный вектор, включающий:

— словарь, содержащий все токены (леммы), встречающиеся в материалах, и их идентификаторы (значение хэш-функции);

— векторы числовых признаков для каждого материала группы, при этом в качестве числового признака использовалась статистическая мера TF-IDF (англ. TF — termfrequency, IDF — inversedocumentfrequency).

HashingVectorizer — функция, преобразующая коллекцию текстовых документов в матрицу scipy.sparse. Данная матрица содержит: значение хэш-функции токена; число вхождений токенов (двоичную информацию о вхождениях: «входит» — 1, «не входит» — 0), если установлен параметр norm = 'l1', а если norm = 'l2', то Евклидово расстояние. Параметры функции HashingVectorizer, заданные по умолчанию, были изменены:

1) параметр n_features = 50000 (по умолчанию *n_features = 1048576*) задает количество признаков, т.е. максимальную длину вектора, с уменьшением значения параметра увеличивается скорость работы векторизатора и уменьшается качество классификатора;

2) параметром stop_words (по умолчанию параметр не задан) задается список стоп-слов для уменьшения зашумленности материала, использовался список стоп-слов, предоставляемый функцией stopwords из библиотеки nltk.

Преимущество использования функции HashingVectorizer заключается в скорости работы и в меньшем объеме используемой оперативной памяти за счет отсутствия необходимости постоянного хранения словаря в памяти. Такой результат достигается посредством использования хэш-функции (32-разрядная версия Murmurhash3).

TfidfTransformer-функция преобразует полученный результат от функции HashingVectorizer (матрица scipy.sparse) в нормализованный вектор с применением статистической меры TF-IDF. Параметры функции TfidfTransformer оставили заданными по умолчанию:

— параметр norm = 'l2' задает метод нормализации, приведенный в формуле (3), предотвращает появление значения веса токена, равного нулю;

— параметр use_idf = True разрешает применить взвешивание IDF;

— параметр smooth_idf = True сглаживает значение IDF, предотвращает деление на ноль;

— параметр sublinear_tf = False, в случае, когда параметр установлен как True, то используется сублинейное масштабирование TF-меры, т.е. применяется 1 + log(tf).

Формула расчета значения TF-IDF, используемая в функции TfidfTransformer, отличается от общепринятой, имеет вид:

$$TF(t,d) = \frac{n_t}{\sum_k n_k}\ , \quad (2)$$

$$IDF(t,D) = \log\left[\frac{1+|D|}{\{d_i \in D|\ t \in d_i\}+1}\right] + 1, \quad (3)$$

$$TF - IDF(t,d,D) = TF(t,d) * IDF(t,D), \quad (4)$$

где n_t есть число вхождений слова t в документ, а в знаменателе общее число слов в данном документе; |D| — число документов в корпусе; $\{di \in D|t \in di\}$ — число документов из коллекции D, в которых встречается слово t (когда $n_t \neq 0$).

По каждому направлению материалов была сформирована обученная модель классификатора с использованием функций RandomForestClassifier и RandomForestClassifier.fit библиотеки scikit-learn.

RandomForestClassifier-функция определяет следующие параметры модели классификатора:

— параметр n_estimators = 500 (по умолчанию n_estimators = 10) задает количество деревьев в лесу, чем больше деревьев, тем выше качество, но время настройки и работы классификатора также пропорционально увеличиваются;

— параметр max_depth = None задает максимальную глубину деревьев, значение None означает без ограничения, при увеличении глубины возрастает качество обучения;

— параметр min_samples_split = 2 устанавливает минимальное количество образцов для разбиения внутреннего узла, при увеличении параметра качество обучения падает.

RandomForestClassifier.fit-функция осуществляет создание и обучение модели классификатора с установленными параметрами на основе нормализованного вектора (результат функции TfidfTransformer) и массива меток (установленные оценки материалов).

Тестирование модели классификатора

Тестирование обученной модели классификатора осуществлялось с использованием функции RandomForestClassifier.predict. Входными параметрами являются обученная модель классификатора и нормализованный вектор материалов группы «контроль», сформированный с использованием функций HashingVectorizer и TfidfTransformer. На выходе был получен одномерный массив с единственным элементом, значением которого является метка (оценка материала) 0 — «разрешен» или 1 — «запрещен».

Полученные результаты тестирования представлены в таблицах 3 (несбалансированная группа «обучение») и 4 (сбалансированная группа «обучение»).

Таблица 3

Несбалансированная группа «обучение»

Направление	Оценка классификатора	Оценка специалиста	
		«запрещенные»	«разрешенные»
Наркоторговля	«запрещенные»	5	20
	«разрешенные»	10	1466
Экстремизм	«запрещенные»	16	66
	«разрешенные»	8	826
Терроризм	«запрещенные»	1	25
	«разрешенные»	3	950

Таблица 4

Сбалансированная группа «обучение»

Направление	Оценка классификатора	Оценка специалиста	
		«запрещенные»	«разрешенные»
Наркоторговля	«запрещенные»	158	108
	«разрешенные»	0	6637
Экстремизм	«запрещенные»	109	54
	«разрешенные»	0	841
Терроризм	«запрещенные»	58	56
	«разрешенные»	0	1850

В таблице 5 представлены полученные результаты тестирования несбалансированной группы «обучение» по коллекции из токенов.

Таблица 5

Несбалансированная группа «обучение» из токенов

Направление	Оценка классификатора	Оценка специалиста	
		«запрещенные»	«разрешенные»
Наркоторговля	«запрещенные»	3	25
	«разрешенные»	3	1534
Экстремизм	«запрещенные»	6	78
	«разрешенные»	4	841
Терроризм	«запрещенные»	0	26
	«разрешенные»	1	959

Оценка качества полученных результатов

Для оценки качества классификатора использовались следующие характеристики: точность (precision показывает, сколько из определенных классификатором материалов как «запрещенные» также были отмечены специалистами), полноты (recall показывает, сколько от общего количества отмеченных материалов как «запрещенные» специалистами также были определены классификатором) и F-меры ($F_{measure}$ объединяет точность и полноту с учетом их значимости или веса):

$$\text{precision} = \frac{TP}{TP+FP}; \quad (5)$$

$$\text{recall} = \frac{TP}{TP+FN}; \quad (6)$$

$$F_{measure} = \frac{1}{\alpha\frac{1}{\text{precision}}+(1-\alpha)\frac{1}{\text{recall}}}, \quad (7)$$

где TP — количество материалов, определенных классификатором как «запрещенные» и специалистом; TN — количество материалов, не определенных как «запрещенные» классификатором и специалистом; FP — количество материалов, определенных классификатором как «запрещенные», но не отмеченных как таковые специалистом; FN — количество материалов, не определенных классификатором как «запрещенные», но отмеченные так специалистом; α — коэффициент, определяющий соотношение весов (значимости) точности и полноты.

Расчеты оценки качества классификатора по результатам обучения на основе несбалансированной обучающей группе материалов, сбалансированной обучающей группе материалов и несбалансированной обучающей группе материалов в виде набора токенов представлены в таблицах 6, 7 и 8 соответственно.

Таблица 6

Оценка качества на несбалансированной группе

Направление	Precision	Recall	F-мера, при $\alpha = 0,3$	F-мера, при $\alpha = 0,5$	F-мера, при $\alpha = 0,7$
Наркоторговля	0,33	0,2	0,227	0,25	0,278
Экстремизм	0,67	0,19	0,248	0,302	0,396
Терроризм	0,25	0,04	0,051	0,067	0,094

Результаты тестирования модели на несбалансированной обучающей группе материалов (табл. 6) неудовлетворительны. Показатели точности и полноты низкие. F-мера характеризует низкое качество классификатора, с незначительным перекосом в сторону точности, так как при увеличении коэффициента α незначительно увеличивается показатель, большее количество материала, помеченное экспертом как «запрещенное», было классифицировано как «разрешенное».

Таблица 7

Оценка качества на сбалансированной группе

Направление	Precision	Recall	F-мера, при $\alpha = 0,3$	F-мера, при $\alpha = 0,5$	F-мера, при $\alpha = 0,7$
Наркоторговля	0,59	1	0,827	0,742	0,673
Экстремизм	0,67	1	0,871	0,802	0,744
Терроризм	0,51	1	0,776	0,675	0,598

Приведенные результаты тестирования модели на сбалансированной обучающей группе материалов (см. табл. 7) являются более чем удовлетворительными. Показатели точности и полноты высокие. F-мера характеризует высокое качество классификатора с перекосом в сторону полноты, так как при увеличении значения коэффициента α уменьшаются показатели. Материалы, оцененные экспертом как «запрещенные», так же точно пометил классификатор.

Таблица 8

Оценка качества на несбалансированной группе из токенов

Направление	Precision	Recall	F-мера, при $\alpha = 0,3$	F-мера, при $\alpha = 0,5$	F-мера, при $\alpha = 0,7$
Наркоторговля	0,5	0,11	0,14	0,176	0,238
Экстремизм	0,6	0,07	0,097	0,128	0,186
Терроризм	0	0	0,0	0,00	0,00

Результаты тестирования модели на несбалансированной обучающей группе материалов в виде токенов (табл. 8) неудовлетворительны. Показатели точности и полноты низкие. F-мера характеризует низкое качество классификатора с незначительным перекосом в сторону точности, так как при увеличении коэффициента α незначительно увеличивается показатель, большее количество материала, помеченное экспертом как «запрещенное», было классифицировано как «разрешенное».

Заключение. Алгоритм машинного обучения «случайный лес» может использоваться для классификации материалов в АИС «Поиск». Для повышения качества его работы нужно обеспечить следующее:

1) увеличить количество документов в обучающей выборке, при этом должны быть представлены все возможные варианты для каждого из классов (репрезентативность выборки);

2) сбалансировать обучающую выборку (количество материалов каждого класса должно быть примерно одинаковым);

3) материалы в обучающей выборке должны содержать значимую текстовую информацию.

Библиографический список

1. Карташев Е.А., Царегородцев А.Л. Автоматизированная информационная система поиска и анализа информации в сети Интернет // Фундаментальные исследования. — 2016. — № 10, ч. 2.

2. Епрев А.С. Автоматическая классификация текстовых документов // Математические структуры и моделирование. — № 21. — 2010.

3. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys. — 34(1) . — 2002.

4. Кафтанников И.Л., Парасич А.В. Об особенности применения деревьев решений в задачах классификации // Вестник ЮУрГУ. Серия: Компьютерные технологии, управление, радиоэлектроника. — 2015. — Т. 15, No 3 [Электронный ресурс]. — URL: https://vestnik.susu.ru/ctcr/article/viewFile/4205/3780.

5. Вьюгин В.В. Математические основы машинного обучения и прогнозирования. — М., 2014.

6. Маннинг Кристофер Д., Рагхаван Прабхакар, Шютце Хайнрих. Введение в информационный поиск. — М., 2014.

7. Терновой О.С., Шатохин А.С. Использование байесовского классификатора для получения обучающих выборок, позволяющих определять вредоносный трафик на коротких интервалах // Известия Алтайского гос. ун-та. — 2013. — №1/1 (77).

8. Терновой О.С. Методика и средства раннего выявления и противодействия угрозам нарушения информационной безопасности в результате ddos атак // Известия Алтайского гос. ун-та. — 2013. — №1/2(77). DOI: 10.14258/izvasu(2013)1.2-24.

9. Андреев А.М., Березкин Д.В., Морозов В.В., Симаков К.В. Автоматическая классификация текстовых документов с использованием нейросетевых алгоритмов и семантического анализа // Мир ПК. — 2007. — № 9.

10. Круглов В.В., Борисов В.В. Искусственные нейронные сети. Теория и практика. — М., 2001.

11. Попков М.И. Автоматическая система классификации текстов для базы знаний предприятия // International Journal of Open Information Technologies : научный журнал. — 2014. Т. 2, No 7 [Электронный ресурс]. — URL:http: // cyberleninka.ru/article/n/avtomaticheskaya-sistema-klassifikatsii-tekstov-dlya-bazy-znaniy-predpriyatiya.

12. Random Forest Classifier [Electronic resourse]. — URL: http://scikit-learn.org/stable/modules/ generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier.

13. Половикова О.Н. Анализ способов формализаций документов для выполнения семантического поиска // Известия Алтайского гос. ун-та. — 2012. — №1 (73).