УДК 004.85

Анализ качества бинарной классификации веб-страниц методом опорных векторов

С.В. Волошин[1], А.Л. Царегородцев[1], Е.А. Карташев[1], В.В. Славский[2]

[1]Югорский научно-исследовательский институт информационных технологий (Ханты-Мансийск, Россия)
[2] Югорский государственный университет (Ханты-Мансийск, Россия)

Support Vector Machines Analysis of Web Pages Binary Classification Quality

S.V. Voloshin[1], A.L. Tsaregorodtsev[1], E.A. Kartashev[1], V.V. Slavskiy[2]

[1]Ugra Research Institute of Information Technologies (Khanty-Mansiysk, Russia)
[2]Ugra State University (Khanty-Mansiysk, Russia)

Представлены результаты анализа качества бинарной классификации веб-страниц методом опорных векторов на наличие информации, распространение которой в Российской Федерации запрещено. Представлены данные для трех коллекций документов: «наркоторговля», «экстремизм», «терроризм». Коллекции документов сформированы по результатам работы специалистов с одной из существующих автоматизированных информационных систем поиска и анализа информации в интернете. Для каждой коллекции описываются соотношения классов, обучающей и тестовой выборок, распределение по типу интернет-ресурсов, а также часть проблем, затрудняющих построение классификатора или понижающих качество классификации. Описывается построение вектора документа. Приводятся полученные результаты тестирования классификатора для различных функций ядра. Для оценки качества используются такие характеристики, как точность, полнота и F1-мера. В качестве реализации метода опорных векторов используется библиотека *scikit-learn*. По результатам классификации тестовой выборки проводится анализ ошибок, делаются заключения о качестве классификации для данных коллекций документов.

Ключевые слова: анализ данных, машинное обучение, метод опорных векторов, классификация текстов, бинарная классификация

This paper presents the analysis of classification quality of web-pages binary classification by the support vector machines method. This classification is required to reveal the web pages containing text information, which distribution is forbidden in Russian Federation. Results are shown for three document collections: “drug dealing”, “extremism” and “terrorism”. Collections of documents are created as a result of specialists’ work with one of the Internet information search and analysis information systems. For each collection, we describe class proportions of testing and training samples, distribution by the type of Internet resources, and several problems that make the classification itself or classifier training difficult. Formation of document’s vector is also described. Next, we show testing results for different kernel functions and analyze classification mistakes. We use precision, recall and F1 score as quality measures. Machine learning library “scikit-learn” is used to implement support vector machines. Finally, we make assumptions about classification quality.

Key words: data analysis, machine learning, support vector machine, text classification, binary classification problem.

DOI 10.14258/izvasu(2017)4-14

Введение. Целью настоящей работы является анализ качества бинарной классификации веб-страниц, размещенных в сети Интернет, на наличие информации, распространение которой запрещено, методом опорных векторов. При достаточном качестве (полнота не менее 0.7, F1-мера не менее 0.65) классификатор предполагается использовать в существующей автоматизированной информационной системе поиска и анализа информации в сети Интернет (далее АИС Поиск). Описание АИС Поиск представле-

но в публикации [1]. Классификатор предполагается использовать в процессе «анализ документов», а результат — при ранжировании выдачи материалов пользователям системы (экспертам) для вынесения экспертной оценки.

Нами был проведен анализ статей на предмет использования метода опорных векторов для классификации текстовой информации. В исследованиях [2–4] приводится математическое описание принципов работы метода опорных векторов. В статье [5] утверждается, что метод опорных векторов применим для классификации текста. Считается [6], что применение регуляризации в методе опорных векторов ведет к робастности метода опорных векторов. В [7] и [8] описываются некоторые особенности библиотеки scikit-learn, которая использовалась в системе для построения классификатора. По результатам анализа можно сделать вывод, что использование метода опорных векторов удовлетворяет особенностям нашей системы. В [9] описаны некоторые способы формального представления документа, в [10–11] приведены примеры использования других алгоритмов классификации, поступающей в систему информации.

Представленная нами работа включает следующие этапы:

1) анализ данных: построение выборки данных, получение эмпирических распределений данных, выявление аномалий и обозначение проблем в данных;

2) построение классификатора: определение признаков и меры признаков, описание структуры словаря, обучение классификатора;

3) оценка качества классификации: сравнение качества классификации при различных настройках, анализ ошибок классификатора;

4) вынесение заключений о качестве классификации.

1. **Анализ данных.** В системе для обучения и тестирования классификатора использовались три коллекции документов: «Наркоторговля», «Экстремизм», «Терроризм».

Под документом подразумевается автоматически очищенное от html-разметки и исполняемого исходного кода текстовое содержимое веб-страницы в интернете, полученной как результат выдачи поисковых машин. Каждый документ в коллекции, который был использован при обучении или тестировании классификатора, имеет одну из двух оценок: «запрещенный» или «разрешенный», что соответствует наличию или отсутствию информации. Количество документов в каждой коллекции представлено в таблице 1.

Таблица 1

Объем коллекций документов

Коллекция	Запрещенные документы	Разрешенные документы	Доля запрещенных материалов в коллекции, %
Наркоторговля	380	8335	4,4
Экстремизм	329	1771	15,7
Терроризм	148	3058	4,6

В каждой коллекции документов имеются преобладающие типы интернет-ресурсов, их распределение значительно различается в каждой тематике (анализ был проведен по 148 документам, помеченным как запрещенные, для каждой коллекции; полученные распределения представлены в таблице 2). Например, в коллекции документов «наркоторговля» среди запрещенных материалов преобладают форумы, а в коллекции «терроризм» — хостинги медиаконтента.

Таблица 2

Распределение документов по типу интернет-ресурса в коллекциях документов

Тип интернет-ресурса	Количество документов в выборке		
	Наркоторговля	Экстремизм	Терроризм
Аудиохостинги	0	12	36
Видеохостинги	5	29	30
Социальные сети	5	38	24
Ресурсы, позиционирующие себя как средства массовой информации	4	23	18
Тематические порталы и блоги	19	26	15
Форумы	37	10	18
Интернет-магазины	76	6	0
Затруднительно определить тип ресурса	2	4	7

В коллекциях документов присутствует ряд проблем, которые затрудняют классификацию:

— Несбалансированность выборок (oversampling): как видно из таблицы 1, в выборке преобладают документы с меткой «разрешенный». Для решения проблемы преобладания в выборке документов, помеченных в системе как разрешенные, была использована система штрафов библиотеки scikit-learn [7].

— Схожесть запрещенного и разрешенного: существуют материалы, содержание которых весьма похоже на материалы, запрещенные к распространению, при этом эксперт оценил материал как разрешенный. Например, продажа семян конопли: такой материал содержит перечисление сортов конопли, свойственный наркоманам сленг и другие признаки, специфичные для ресурсов, связанных с наркоторговлей. Однако продажа семян законом не запрещена, поэтому эксперт оценивает материал как разрешенный. Данная особенность характерна для всех коллекций документов. В работе Huan Xu [5] утверждается, что влияние таких документов на результат классификации можно снизить, применив регуляризацию. В данной работе использовалась L_2-регуляризация.

— Веб-страницы, содержащие нетекстовую запрещенную информацию, например видео- и аудиохостинги.

— Слишком малый или слишком большой объем веб-страницы: тексты веб-страниц не имеют фиксированного объема. Количество словоформ в двух взятых случайно веб-страницах может различаться на порядки.

2. **Построение классификатора.** Для обучения классификатора нужны векторы документов и метки классов. Метками классов являются оценки экспертов. Вектор документа представляет собой последовательность пар вида {идентификатор признака: мера признака} с соответствующей последовательности меткой класса. В качестве признака выступает порядковый номер словоформы в словаре.

Словарь строит вектор документа по всем его лексемам. Он содержит лексемы, составленные только из латинских и кириллических символов, без цифр, иероглифов, рун и т.п. Каждая лексема в словаре присутствует в единственном числе. Из сложных словоформ (например, ‘Абу-Даби’) также выделяется лексема. Перед прохождением проверки на наличие лексемы словоформы в словаре дополнительно выполняются другие преобразования (сведение к нижнему регистру, удаление диакритических знаков). Словоформы, не являющиеся словом (тег, число, url, email), отбрасываются. Лексемы сложных словоформ определяются как одна лексема. Размер словаря составил чуть более миллиона признаков. В качестве меры признака было взято количество вхождений лексемы (признака) в документ.

Коллекции документов были разделены на обучающую и тестовую выборки (табл. 3).

Таблица 3

Объемы обучающих и тестовых выборок

Коллекция	Оценка	Объем обучающей выборки (в документах)	Объем тестовой выборки (в документах)	Доля тестовой выборки, %
Наркоторговля		6100	2615	30
Наркоторговля	Запрещенные	274	106	28
Наркоторговля	Разрешенные	5826	2509	30
Экстремизм		1448	652	31
Экстремизм	Запрещенные	209	120	37
Экстремизм	Разрешенные	1239	532	30
Терроризм		2222	984	31
Терроризм	Запрещенные	99	49	33
Терроризм	Разрешенные	2123	935	31

3. **Описание полученных результатов.** В процессе апробации классификатора было опробовано несколько функций ядра: линейная, полиномиальная, сигмоидальная и радиального базиса (RBF). На настройках по умолчанию наилучший результат классификации (в соответствии с мерой F1) был получен при использовании функции радиального базиса как функции ядра. Полученные результаты классификации (по классу «запрещенный») при дополнительной настройке классификаторов представлены в таблице 4. Параметры для дополнительной настройки получены путем «поиска по сетке», (описание представлено в работе [12]).

Таблица 4

Результаты классификации с применением разных функций ядра

Ядро	Тематика	Полнота	Точность	F1 мера
Линейная	Наркоторговля	0.72	0.67	0.69
	Экстремизм	0.50	0.67	0.57
	Терроризм	0.55	0.52	0.54
RBF	Наркоторговля	0.71	0.81	0.75
	Экстремизм	0.58	0.68	0.62
	Терроризм	0.51	0.68	0.58
Полиномиальная	Наркоторговля	0.76	0.64	0.69
	Экстремизм	0.62	0.66	0.64
	Терроризм	0.51	0.60	0.55
Сигмоидальная	Наркоторговля	0.75	0.63	0.68
	Экстремизм	0.60	0.73	0.66
	Терроризм	0.51	0.60	0.55

Анализ выбранных случайно документов из коллекции «наркоторговля», при определении которых ошибся классификатор, представлены в таблице 5.

Исходя из полученных результатов можно сделать вывод, что для коллекции документов «наркоторговля» большинство ошибок классификатора поддается анализу, и в дальнейшем возможно устранить часть таких ошибок.

Для коллекций документов «терроризм» и «экстремизм» провести анализ ошибок затруднительно. Очевидно, что классификатор теряет большое количество важной информации из-за того, что наиболее значимая информация в документах может содержаться в медиафайлах.

Таблица 5

Описание ошибок классификатора

Класс ошибки	Описание материала	Доля ошибок, %
Документ ошибочно помечен как разрешенный	Материал содержит малое количество значимой информации либо большое количество нерелевантной информации	39
	Текст материала недоступен системе	15
	Описание наркотического вещества завуалировано либо нетипично	15
	Явная ошибка классификатора	31
Документ ошибочно помечен как запрещенный	Страницы с сайтов аптек, медицинских форумов и т.д., где находится подробное описание наркотических веществ или лекарственных средств	22
	Описание воздействия наркотических веществ или сильных обезболивающих на организм и умственную деятельность человека	13
	Продажа семян конопли, галлюциногенных грибов и т.п.	7
	Новостные публикации, содержащие информацию о наркотических веществах	11%
	Словари жаргона наркоманов	4
	Материал не на русском языке	17
	Материал содержит видеоконтент	2
	Материал содержит обсуждение анонимности и приватности на ресурсах интернета, связанных с наркоторговлей	4
	Материал перегружен автоматически сгенерированным текстом или большим количеством несвязной информации	11
	Явная ошибка классификатора	9

Эмпирический анализ результатов классификации тестовой выборки документов в данных коллекциях показывает, что подавляющее большинство документов в этой выборке, оцененных классификатором как запрещенные, имеют в своем содержимом имена известных пропагандистов, а также слова, характерные пропаганде ислама. В то же время документы, помеченные классифи-

катором как разрешенные, таких слов и имён практически не содержат.

Заключение. Исходя из показателей полноты и точности классификации, а также распределения ошибок классификатора можно сделать вывод, что построение классификатора для коллекции документов «наркоторговля» можно считать обоснованным.

По причинам, описанным выше, использование метода опорных векторов при классификации коллекций документов «терроризм» и «экстремизм» затруднительно. Можно высказать предположение, что классификация по ключевым словам на данный момент будет эффективней для данных коллекций документов (например, в работе [13] описывается обучаемый классификатор веб-сайтов по ключевым словам и приведен ряд предложений по улучшению качества классификации). Также может иметь смысл повторная апробация классификатора для коллекций документов «терроризм» и «экстремизм» после увеличения количества документов класса «запрещенный».

Библиографический список

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

2. Вьюгин В.В. Математические основы машинного обучения и прогнозирования [Электронный ресурс]. — URL: http://e.lanbook.com/book/56397.

3. Fradkin D., Muchnik I. Support Vector Machines for Classification// Abello J. Carmode G. (Eds); Discrete Methods in Epidemiology, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, volume 70, 2006.

4. Cristianini N., Shawe-Taylor J. An Introduction to Support Vector Machines and other kernel-based learning methods. — Cambridge, 2000.

5. Joachims T. Text categorization with support vector machines: learning with many relevant features // Proceedings of ECML-98, 10th European Conference on Machine Learning, Chemnitz, 1998.

6. Xu H., Caramanis C., Mannor Sh. Robustness and Regularization of Support Vector Machines // The Journal of Machine Learning Research, 10, 12/1/2009.

7. Support vector machines: scikit-learn [Electronic resourse]. — URL: http://scikit-learn.org/stable/modules/svm.html (дата обращения: 01.03.17).

8. Unbalanced problems of support vector machines: scikit-learn [Electronic resourse]. URL: http://scikit-learn.org/stable/modules/svm.html#unbalanced-problems (дата обращения: 01.03.17).

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

10. 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.

11. 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).

12. Ямшанов М.Л. Оптимизация выбора параметров SVM-классификатора с ядром RBF для задач классификации текстовых документов // Вестник ВятГГУ. — 2006. — №15.

13. Маслов М.Ю., Пяллинг А.А., Трифонов С.И. Автоматическая классификация веб-сайтов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции : труды Десятой Всерос. науч. конф. «RCDL'2008». — Дубна, 2008.