МЕТОДИКА МНОГОАСПЕКТНОЙ ОЦЕНКИ И КАТЕГОРИЗАЦИИ ВРЕДОНОСНЫХ ИНФОРМАЦИОННЫХ ОБЪЕКТОВ В СЕТИ ИНТЕРНЕТ

А.А. Браницкий[1*], **И.Б. Саенко**[1]

[1]Санкт-Петербургский институт информатики и автоматизации Российской академии наук, Санкт-Петербург, 199178, Российская Федерация
*Адрес для переписки: alexander.branitskiy@gmail.com

Информация о статье
УДК 004.056
Статья поступила в редакцию 30.04.2019

Ссылка для цитирования: Браницкий А.А., Саенко И.Б. Методика многоаспектной оценки и категоризации вредоносных информационных объектов в сети Интернет // Труды учебных заведений связи. 2019. Т. 5. № 3. С. 58–65. DOI:10.31854/1813-324X-2019-5-3-58-65

Аннотация: *В условиях быстрого развития информационных технологий возникает задача, связанная с обнаружением источников вредоносной информации в сети Интернет. Для ее решения могут применяться методы машинного обучения как один из наиболее популярных и мощных инструментов, предназначенных для выявления зависимостей между входными (наблюдаемыми) данными и выходными (желаемыми) результатами. В данной статье представлена методика, направленная на многоуровневую обработку входных данных о вредоносных информационных объектах в сети Интернет и обеспечивающая их многоаспектную оценку и категоризацию с использованием методов машинного обучения. Цель исследования заключается в повышении эффективности процесса обнаружения вредоносной информации в сети Интернет на примере задачи классификации веб-страниц.*

Ключевые слова: *информационные объекты, вредоносная информация, классификаторы, веб-страницы, многоуровневая схема комбинирования.*

Введение

Современная глобальная сеть содержит огромное количество разнородной информации, которая по своему содержанию может рассматриваться как вредоносная. Обнаружение источников является важной задачей, поскольку их распространение и использование может приводить к серьезным негативным последствиям как на локальном уровне, затрагивающем интересы и права отдельных лиц, так и на глобальном уровне, находящем отражение в международных разногласиях и конфликтах.

В качестве примера обнаружения вредоносных информационных объектов (ИО), можно привести системы родительского контроля. В роли ИО в таких системах выступает информация, предоставляемая такими Интернет-сервисами, как веб-сайты, социальные сети, онлайн-чаты, игровые и медиапорталы и другие. В этом случае запрещение доступа к ИО выполняется на основе анализа потоков данных, передаваемых от соответствующего Интернет-ресурса к конечному пользователю. Признаком для такого запрета может являться наличие вредоносной информации, содержащей, например, ненормативную лексику, призывы к противоправным действиям или указания, пропагандирующие нездоровый образ жизни.

Саму задачу обнаружения вредоносной информации можно рассматривать как задачу категоризации ИО, в которой заранее определены нелегитимные категории. Системы, предназначенные для решения этой задачи, могут быть основаны как на ручном построении классификационных правил, так и с привлечением автоматических средств их генерации. Именно последний тип подобных систем представляет наибольший интерес со стороны исследователей в связи с постоянным ростом, развитием и популяризацией такого перспективного научного направления, как машинное обучение.

В статье рассматривается вопрос повышения показателей эффективности обнаружения вредоносных ИО на примере задачи классификации веб-страниц с использованием различных методов машинного обучения и их комбинирования.

Релевантные работы

Задача обнаружения вредоносной информации в сети Интернет может быть сведена к классификации веб-страниц, в которой ряд категорий заранее определен системным администратором как содержащий нелегитимный контент. В этой области существует множество работ, посвященных построению, как экспертных систем, так и полностью автоматических систем.

Представленная в [1] система CONSTRUE основана на продукционных правилах, создаваемых вручную оператором-экспертом. Данная система предназначена для классификации экономических и финансовых новостей и соотнесения анализируемого текста к одной из 674 категорий. Точность классификации для системы CONSTRUE составляет более 90 %. Недостаток такой системы заключается в том, что ее поддержание в консистентном состоянии требует регулярного привлечения специалистов, выполняющих добавление и корректирование продукционных правил.

Подход к категоризации содержимого с автоматической генерацией правил классификации рассматривается исследователями C. Apté, F. Damerau и S.M. Weiss [2]. Предлагаемый ими формат правил – дизъюнктивная нормальная форма (ДНФ). Алгоритм формирования правил основан на последовательной замене одного из конъюнктов и дальнейшем добавлении нового конъюнкта до тех пор, пока не будет построено стопроцентное покрытие обучающей выборки (т. е. такой набор правил, которые будут обеспечивать безошибочную классификацию обучающих элементов). Данный алгоритм выполняет эвристический поиск таких правил: алгоритм не обеспечивает нахождение минимальной по количеству конъюнктов ДНФ. Кроме того, в отличие от дерева решений конъюнкты, объединенные одним правилом при помощи этого алгоритма, не являются взаимоисключающими.

В качестве элементарных конъюнктов (атомов) использовались предикаты, отражающие признаки:

1) вхождения определенного слова (или словосочетания) из локального словаря (набора слов, содержащего специфичные для одной категории понятия) в анализируемый текст;

2) превышения частоты встречаемости определенного выражения внутри анализируемого текста на указанное пороговое значение.

Предложенный подход позволяет сохранить представление правил в формате, удобном для анализа экспертами. В то же время при описанном способе генерации правил теряются обобщающая способность системы и способность обработки зашумленных данных.

Авторы статьи [3] предлагают принять анализируемый документ как массив вещественнозначных коэффициентов, которые представляют собой относительные и абсолютные частоты вхождения определенных слов в классифицируемый текст. Среди таких коэффициентов были выделены:

– частота слова (TF, *от англ.* Term Frequency);

– обратная частота документа (IDF, *от англ.* Inverse Document Frequency);

– важность слова (TD, *от англ.* Term Discrimination), где TD = TF × IDF;

и некоторые другие.

В [4] изложен подход, который позволяет приписывать каждому слову его интегральный вес, включающий вероятность появления этого слова, как в рамках определенной категории, так и внутри всей коллекции документов, и с учетом остальных категорий.

Сравнение двух методов машинного обучения, а именно байесовского классификатора и дерева решений, в рамках задачи категоризации текста выполнено в [5]. Авторы этой статьи подчеркивают, что на крупных наборах обучающих данных лучшую производительность демонстрирует дерево решений, а на более мелких наборах данных – байесовский классификатор. Причем для байесовского классификатора с увеличением числа обрабатываемых признаков наблюдается ситуация переобучения (на контрольном множестве производительность классификатора снижается), а для дерева решений при этих же условиях и достаточном объеме обучающей выборки происходит увеличение показателей эффективности классификации.

Применимость другого популярного метода машинного обучения, а именно машины опорных векторов (МОВ), к задаче классификации текстов исследуется в статье T. Joachims [6], где автор выделяет способность МОВ обучаться как на высокоразмерных, так и на разреженных векторах признаков. Решение задачи классификации текстов в большинстве случаев имеет вид линейно разделимых областей, для обособления которых может использоваться МОВ.

В статье R. Johnson и T. Zhang [7] описываются два типа конволюционных нейронных сетей: прямое распространение сигнала и с преобразованием «мешка» слов (bag-of-words) на конволюционном слое. В результате экспериментов авторами было выявлено, что первый тип нейронной сети демонстрирует бóльшую производительность в терминах показателей классификации по сравнению со вторым типом нейронной сети.

В [8] представлен метод для извлечения признаков в рамках задачи категоризации текста. Предложенная модификация генетического алгоритма, как показывают эксперименты, позволяет добиться более компактного представления обучающих векторов в терминах их размерности и повысить качество классификации анализируемого текста.

Общим ограничением для вышеупомянутых работ, посвященных приложению методов машинного обучения к решаемой задаче, является приме-

нение односоставных классификаторов, что приводит к невозможности обучения модели по частям и, в свою очередь, затрудняет возможность распараллеливания этого процесса.

Анализ работ в данной предметной области показывает актуальность рассматриваемой темы. В то же время, несмотря на их разнообразие, задача разработки методики, предназначенной для обнаружения вредоносных ИО в сети Интернет и сочетающей в себе методы машинного обучения и их комбинирования, остается по-прежнему высокоприоритетным направлением в научно-исследовательском сообществе.

Многоаспектная оценка и категоризация вредоносных информационных объектов в сети Интернет

В разработанной авторами методике выделяется пять шагов (рисунок 1).

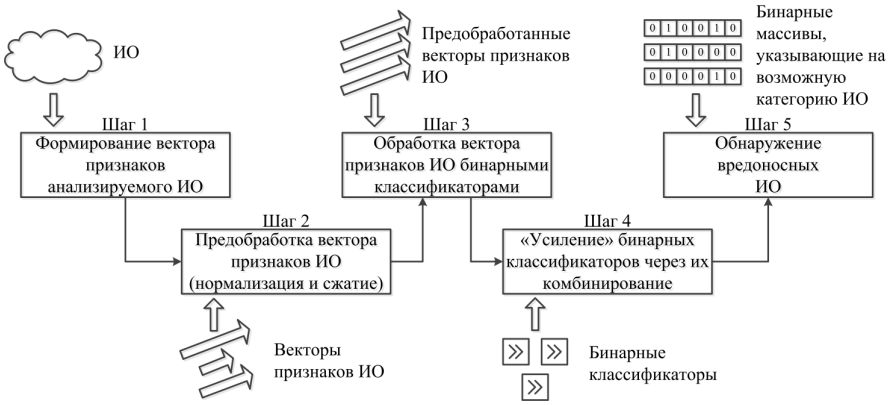

Рис. 1. Пошаговая схема методики многоаспектной оценки и категоризации вредоносных ИО в сети Интернет

На *первом шаге* осуществляется формирование вектора признаков ИО. *Второй шаг* направлен на выполнение процедуры предобработки созданного вектора признаков при помощи его покомпонентной нормализации и уменьшения размерности. На *третьем шаге* вектор признаков анализируемого ИО обрабатывается бинарными классификаторами. *Четвертый шаг*, «усиление» бинарных классификаторов, характеризуется их комбинированием в единый высокоуровневый классификатор, который предоставляет данные, необходимые на *пятом шаге* для формирования окончательного решения о наличии вредоносного содержимого внутри ИО. Вредоносность содержимого ИО задается через принадлежность ИО к одной из тех категорий, которые заранее были определены оператором системы как содержащие нелегитимную информацию.

На рисунке 2 представлена многоуровневая схема комбинирования классификаторов, покрывающая шаги 2, 3 и 4 в системе многоаспектной оценки и категоризации вредоносных ИО. Шаги 1 и 5 рассмотрены ниже в разделе «Эксперименты».

В рамках этой схемы выделяется три уровня:

1) предобработка входного объекта при помощи минимаксной (min-max) нормализации и метода главных компонент (МГК);

2) обработка входного объекта при помощи разнообразных базовых классификаторов, построенных на основе методов машинного обучения;

3) агрегирование базовых классификаторов через их композицию, представленную как метод взвешенного голосования (МВГ).

Первый уровень выполняет роль начальной подготовки обучающих и тестовых данных, подаваемых на вход базовых классификаторов. Каждый компонент обрабатываемого вектора масштабируется до неотрицательного значения, не превосходящего 1. Затем извлекаются наиболее информативные признаки. Они получаются как линейная комбинация обрабатываемого вектора и собственных векторов матрицы ковариации, сформированной на основе обучающего набора данных.

Второй уровень содержит пять базовых классификаторов: МОВ, метод k-ближайших соседей (МБС), наивный байесовский классификатор (НБК), линейную регрессию (ЛР) и дерево решений (ДР). Каждый из этих классификаторов является бинарным, т. е. предназначен для обособления объектов только одного фиксированного класса от других. Посредством использования таких классификаторов становится возможным легко и эффективно выполнять их параллельное обучение.

Предположим, что обучающий набор данных содержит m классов и представляется как набор пар:

$$(\boldsymbol{X}, \Omega) = \{(\boldsymbol{x_k}, c_k)\}_{k=1}^{M},$$

где $\boldsymbol{x_k} = (x_{k1}, \dots, x_{kn})^T \in \boldsymbol{X}$ – вектор признаков; c_k – присвоенная вектору $\boldsymbol{x_k}$ метка класса, такая, что $\exists \omega_j \in \Omega = \{\omega_1, \dots, \omega_m\}: \boldsymbol{x_k} \in \omega_j \wedge j = c_k$.

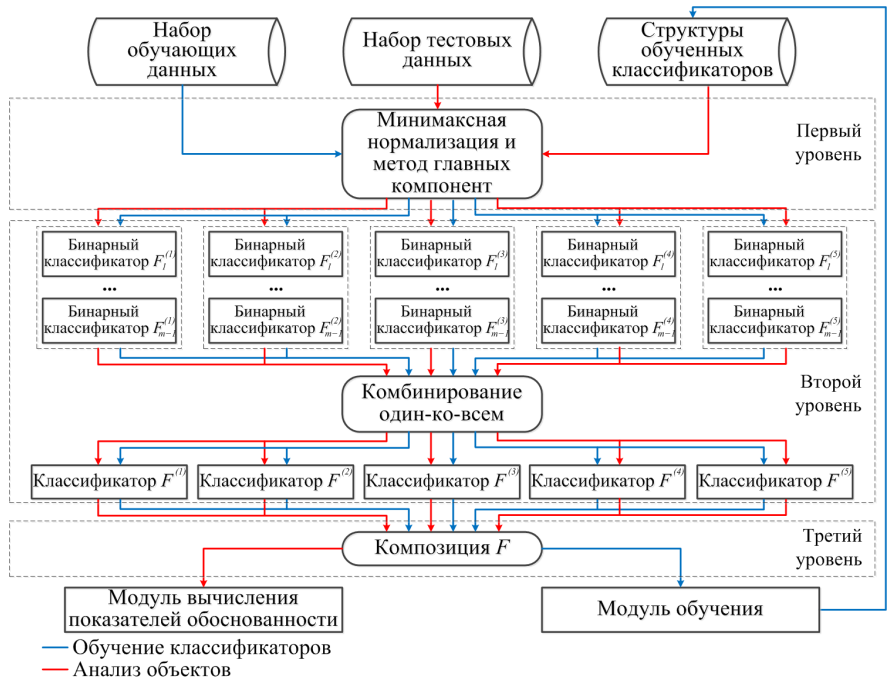

Рис. 2. Многоуровневая схема комбинирования классификаторов в системе многоаспектной оценки и категоризации вредоносных ИО в сети Интернет

В рамках схемы «один-ко-многим» бинарный классификатор $F_j^{(i)}: R^n \to \{0,1\}, (j = 1, \dots, m-1, m \geq 2; i = 1, \dots, 5)$, являющийся частью составного классификатора $F^{(i)}: R^n \to 2^{\{1,\dots,m\}}$, обучается на элементах $\{(\boldsymbol{x_k}, I(c_k = j))\}_{k=1}^{M}$, где I обозначает логическую функцию.

Функционирование составного классификатора $F^{(i)}$ может быть описано следующим образом:

$$F^{(i)}(\boldsymbol{z}) = \begin{cases} \{m\}, \text{если } \forall j \in \{1, \dots, m-1\}\, F_j^{(i)}(\boldsymbol{z}) = 0 \\ \left\{j \middle| F_j^{(i)}(\boldsymbol{z}) = 1\right\}_{j=1}^{m-1}, \text{иначе.} \end{cases}$$

В данной формуле первые $(m-1)$ бинарных классификаторов подвергаются объединению. Бинарные классификаторы $F_j^{(i)}$ настраиваются таким образом, чтобы их выходы равнялись 1, если индекс j соответствует метке класса c_k.

Объект $\boldsymbol{z}$ распознается как принадлежащий классу, промаркированному меткой m, если выходы всех $(m-1)$ бинарных классификаторов установлены в 0. В противном случае выход составного классификатора $F^{(i)}$ представляет собой множество тех меток классов, которые соответствуют индексам бинарных классификаторов с ненулевыми выходами.

Выбор схемы комбинирования «один-ко-многим» обусловлен тем фактом, что она направлена на объединение классификаторов, число которых прямо пропорционально количеству классов. В отличие от других популярных схем [9], таких как «один-к-одному» или направленный ациклический граф, которые обладают квадратичной зависимостью между указанными параметрами, схема «один-ко-многим» является особенно выигрышной в случае нескольких десятков классов.

Третий уровень – это композиция F, которая выражается как МВГ. Настроенный в процессе обучения этой композиции коэффициент w_i отражает вклад соответствующего составного классификатора $F^{(i)}$ в финальное решение о принадлежности анализируемого ИО к тому или иному классу.

Для систем, обеспечивающих поддержку предложенной методики, выделяется два режима: обучение классификаторов и анализ объектов. Первый начинается с загрузки обучающих данных и заканчивается сериализацией структур обученных классификаторов. Второй начинается с загрузки тестовых данных и структур обученных классификаторов и заканчивается вычислением показателей эффективности классификаторов. На рисунке 2 эти режимы обозначены синими (обучение классификаторов) и красными (анализ объектов) стрелками.

Эксперименты

В таблице 1 представлены размеры исследуемого набора данных [10] для каждой из 19 категории. Суммарная мощность исследуемого набора составляет 74893 записи, которые предварительно были разбиты по 19 категориям. Некоторые из них содержат вредоносный контент (например, marijuana, adult, violence) с точки зрения содержания. Сами записи представляют собой веб-страницы, доступные в сети Интернет (см. таблицу 2).

ТАБЛИЦА 1. Категории экспериментального набора и их мощность

№	Категория	Количество
1.	Adult / для взрослых	6748
2.	Alcohol / спиртные напитки	2802
…	…	…
9.	Jew Related / антисемитизм	3446
10.	Marijuana / марихуана	5365
…	…	…
17.	Violence / насилие	1892
18.	Weapon / оружие	2448
…	…	…

Формируемый вектор признаков содержит 402 компонента. При их вычислении использовались три типа исходных данных: URL-строка (адрес Интернет-ресурса), структура документа (статистические данные по HTML-тегам), извлеченный текст (полученный после удаления HTML-тегов).

В качестве показателей эффективности применялись следующие параметры: точность (PR, *от англ.* Precision,); полнота (RC, *от англ.* Recall); *F*-мера (FM, *от англ.* F-measure); аккуратность (AC, *от англ.* Accuracy). На рисунке 3 представлены значения этих показателей, вычисленные для каждого классификатора с использованием данных, не встречавшихся в процессе обучения. Для обучения использовалась выборка, размером приблизительно ¼ от всего набора данных. За счет комбинирования классификаторов при помощи МВГ показатель AC вырос более, чем на 1,5 % по сравнению с наибольшим значением аналогичного показателя, демонстрируемым среди базовых классификаторов, а именно ЛР. Кроме того, значения остальных показателей также увеличились: PR – на 2,63 %; RC – на 2,66 %; FM – на 2,65 %.

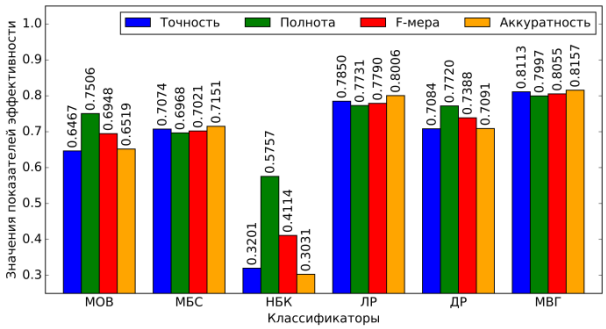

Рис. 3. Показатели эффективности, вычисленные на тестовом наборе данных

На рисунке 4 представлены временные характеристики процесса обучения для каждого классификатора и для различных размерностей входного вектора, полученного в результате сжатия при помощи МГК.

ТАБЛИЦА 2. Параметры веб-страницы

Номер	Описание	Тип исходных данных
1	Абсолютное число выбранных 15 тегов	Структура документа
2–16	Относительное число выбранных 15 тегов	Структура документа
17	Абсолютное число всех тегов	Структура документа
18	Абсолютное число атрибутов во всех тегах	Структура документа
19	Количество ссылок	Структура документа
20	Размер заголовка	Извлеченный текст
21	Размер отображаемого текста	Извлеченный текст
22	Длина комментариев	Структура документа
23–41	Логический признак вхождения имени каждой категории в заголовок	Извлеченный текст
42–60	Логический признак вхождения имени каждой категории в отображаемый текст	Извлеченный текст
61–250	Абсолютная частота вхождения каждого из 10 наиболее употребительных слов, характерных для каждой категории, в отображаемый текст	Извлеченный текст
251–269	Абсолютная суммарная частота вхождения 10 наиболее употребительных слов, характерных для каждой категории, в заголовок	Извлеченный текст
270–288	Абсолютная суммарная частота вхождения 10 наиболее употребительных слов, характерных для каждой категории, в URL	Извлеченный текст и URL-строка
289–345	Степень семантической схожести имени каждой категории с тремя наиболее употребительными словами, входящими в отображаемый текст анализируемого документа, в терминах word2vec [11]	Извлеченный текст
346–402	Степень семантической схожести словаря, состоящего из пяти наиболее употребительных слов в рамках каждой категории, с тремя наиболее употребительными словами, входящими в текст анализируемого документа, в терминах word2vec	Извлеченный текст

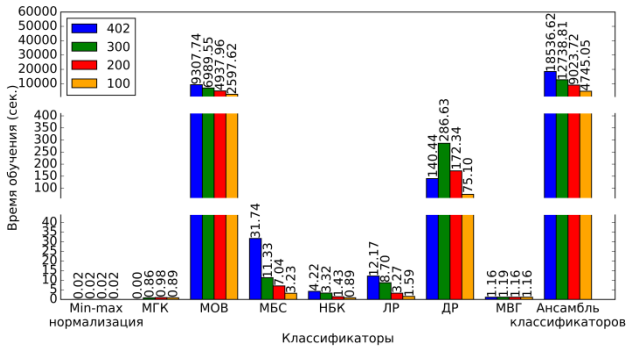

Рис. 4. Временные характеристики процесса обучения

Среди базовых классификаторов наименьшими временными затратами по обучению обладает НБК, а МОВ характеризуется наиболее длительным процессом обучения. Отметим, что изменение размерности признакового пространства прямо пропорционально влияет и на время обучения композиции классификаторов. Так, при сужении признакового пространства с 402 до 300 компонентов наблюдается уменьшение времени обучения ансамбля классификаторов в 1,46 раза, при сужении до 200 компонентов – в 2,05 раза, а до 100 компонентов – в 3,90 раза.

На рисунке 5 представлена зависимость времени, затраченного на анализ 74 893 векторов признаков при помощи каждого классификатора, от размерности обрабатываемого вектора.

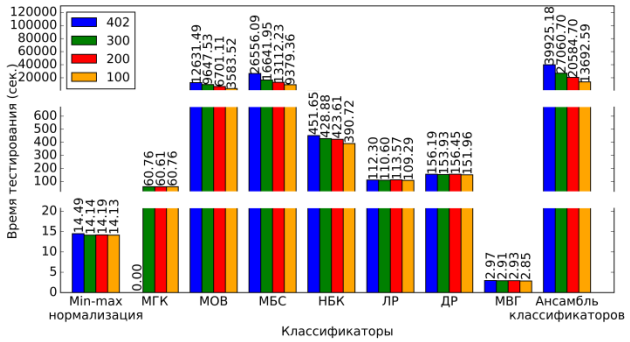

Рис. 5. Временные характеристики процесса анализа

По скорости анализа векторов среди базовых классификаторов наилучшей характеристикой обладает ЛР, на долю которой приходится от 2,8 % до 8 % времени от общего времени функционирования всего ансамбля классификаторов. Напротив, МБС является наиболее трудоемким и требует от 61,5 % до 68,5 % суммарного времени, расходуемого коллективом классификаторов (параметр k в МБС был выбран равным 10). Переход от 402-мерного к 100-мерному вектору признаков, анализируемому каждым базовым классификатором, привел к сокращению времени обработки (формирования результата классификации) этого вектора более чем в 2,9 раза. Таким образом, четырехкратное сжатие признакового пространства позволило сократить время обучения иерархического классификатора и время анализа входного вектора с его помощью почти в 4 и в 3 раза, соответственно.

На рисунке 6 представлена зависимость показателей эффективности, вычисленных для ансамбля классификаторов (композиции на основе МВГ), от размерности входного вектора признаков ИО.

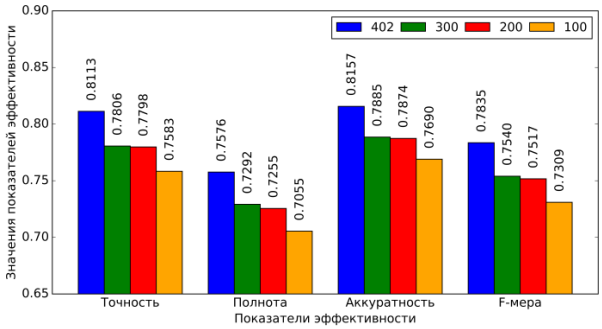

Рис. 6. Показатели эффективности, вычисленные на на основе МВГ

Как и следовало ожидать, с уменьшением размерности анализируемого вектора качество классификации по всем показателям уменьшается. Это обусловливается тем, что при переходе к суженному набору признаков теряется часть информации об исходных данных.

Для вычисления несмещенной оценки выбранных показателей эффективности использовалась пятиблочная кросс-валидация. Одновременно с этим размер обучающей выборки увеличился более чем в 3 раза, что позволило обеспечить лучшее покрытие тестового множества по сравнению со случаем, представленным на рисунке 3. На рисунке 7 представлены усредненные значения этих показателей вместе с отклонениями, обозначенными в виде вертикальных линий, пронизывающих каждый столбик гистограммы. Как видно из рисунка, наилучшими показателями эффективности обладает МВГ.

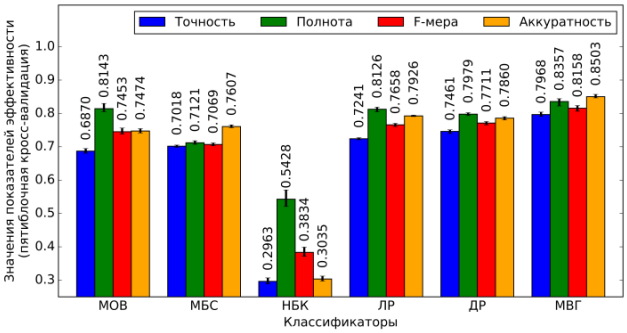

Рис. 7. Показатели эффективности, вычисленные при помощи пятиблочной кросс-валидации

На рисунке 8 показана зависимость времени обучения классификаторов и их ансамбля от числа используемых потоков. Сжатие векторов признаков осуществлялось до 100 компонент. В случае трех классификаторов (МОВ, МБС и ДР) наблюдается заметное снижение времени обучения с увеличением числа задействованных потоков.

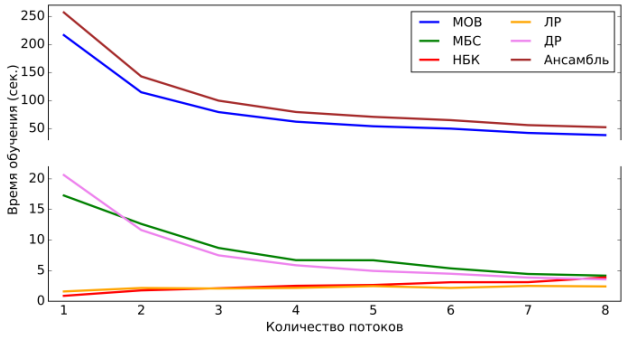

Рис. 8. Зависимость времени обучения базовых классификаторов и их коллектива от количества потоков

Для «быстро обучаемых» классификаторов (НБК и ЛР) такого существенного выигрыша нет. Это обосновывается тем, что прием распараллеливания является эффективным только тогда, когда время инициализации и планирования самого потока является несопоставимо малым по сравнению со временем выполнения самой задачи, запущенной в рамках созданного потока. Суммарное время обучения коллектива из 5 базовых классификаторов с использованием 8 потоков сократилось более чем в 4,9 раза по сравнению с однопоточным режимом.

Заключение

Представленная в статье методика отличается от существующих наличием многоуровневой схемы комбинирования бинарных классификаторов, а также сочетает в себе механизмы ускорения процесса обучения за счет распараллеливания и сжатия признакового пространства.

Многоаспектная категоризация обеспечивается за счет разностороннего анализа веб-документа согласно трем типам исходных данных, а именно URL-строки, структуры документа и извлеченного текста. Многоаспектность оценки обеспечивается за счет проведения разнообразных экспериментов, покрывающих как вычисление показателей эффективности, так и временных показателей, затраченных в процессе обучения классификаторов и анализа объектов. Показатель аккуратности был увеличен более чем на 1,5 % по сравнению с аналогичным показателем, демонстрируемым лучшим базовым классификатором, а именно ЛР. Использование 8 потоков привело к сокращению времени обучения коллектива классификаторов почти в 5 раз по сравнению однопоточным режимом. По скорости анализа векторов среди базовых классификаторов наилучшей характеристикой обладает ЛР. Напротив, МБС является наиболее трудоемким.

БЛАГОДАРНОСТИ

Работа выполнена при финансовой поддержке проекта РНФ № 18-11-00302 в СПИИРАН.

Список используемых источников

1. Hayes P.J., Andersen P.M., Nirenburg I.B., Schmandt L.M. TCS: a shell for content-based text categorization // Proceedings of the Sixth Conference on Artificial Intelligence Applications (Santa Barbara, USA, 5–9 May 1990). Piscataway, NJ: IEEE, 1990. Vol. 1. PP. 320–326. DOI:10.1109/CAIA.1990.89206

2. Apté C., Damerau F., Weiss S.M. Automated learning of decision rules for text categorization // ACM Transactions on Information Systems (TOIS). 1994. Vol. 12. Iss. 3. PP. 233–251. DOI:10.1145/183422.183423

3. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information Processing & Management. 1988. Vol. 24. Iss. 5. PP. 513–523. DOI:10.1016/0306-4573(88)90021-0

4. Fattah M.A. A Novel Statistical Feature Selection Approach for Text Categorization // Journal of Information Processing Systems. 2017. Vol. 13. Iss. 5. PP. 1397–1409.

5. Lewis D.D., Ringuette M. A Comparison of Two Learning Algorithms for Text Categorization // In: Third Annual Symposium on Document Analysis and Information Retrieval. 1994. PP. 81–93.

6. Joachims T. Text categorization with Support Vector Machines: learning with many relevant features // Proceedings of the 10th European Conference on Machine Learning (ECML, Chemnitz, Germany, 21–23 April 1998). Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Berlin, Heidelberg: Springer, 1998. Vol. 1398. PP. 137–142. DOI:10.1007/BFb0026683

7. Johnson R., Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks // Proceeding of the Annual Conference of the North American Chapter of the Association for Computational Linguistics "Human Language Technologies" (Denver, USA, 31 May – 5 June 2015). Stroudsburg: Association for Computational Linguistics, 2015. PP. 103–112. DOI:10.3115/v1/N15-1011

8. Ghareb A.S., Bakar A.A., Hamdan A.R. Hybrid feature selection based on enhanced genetic algorithm for text categorization // Expert Systems with Applications. 2016. Vol. 49. Iss. C. PP. 31–47. DOI:10.1016/j.eswa.2015.12.004

9. Lorena A.C., De Carvalho A.C., Gama J.M.P. A review on the combination of binary classifiers in multiclass problems // Artificial Intelligence Review. 2008. Vol. 30. Iss. 1–4. DOI:10.1007/s10462-009-9114-9

10. Kotenko I., Chechulin A., Shorov A., Komashinsky D. Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking // Proceeding of the 14th Industrial Conference on Data Mining "Advances in Data Mining. Applications and Theoretical Aspects" (ICDM, St. Petersburg, Russia, 16–20 July 2014). Lecture Notes in Computer Science. Cham: Springer, 2014. Vol. 8557. PP. 39–54. DOI:10.1007/978-3-319-08976-8_4

11. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space. 2013. URL: https://arxiv.org/pdf/1301.3781 (дата обращения 10.04.2019)

* * *

THE TECHNIQUE OF MULTI-ASPECT EVALUATION AND CATEGORIZATION OF MALICIOUS INFORMATION OBJECTS ON THE INTERNET

A. Branitskiy[1], **I. Saenko**[1]

[1]Saint-Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences,
St. Petersburg, 193232, Russian Federation

Article info
The article was received 30 April 2019

For citation: Branitskiy A., Saenko I. The Technique of Multi-aspect Evaluation and Categorization of Malicious Information Objects on the Internet. *Proceedings of Telecommunication Universities*. 2019;5(3):58–65. (in Russ.) Available from: https://doi.org/10.31854/1813-324X-2019-5-3-58-65

Abstract: *Under the influence of rapid development in the sphere of information technologies, rises the challenge related to detection of malicious information sources on the Internet. To solve this we can use machine learning methods as one of the most popular and powerful tools designed to identify dependencies between input (observed) data and output (desired) results. This article presents a methodology which is aimed at multi-level processing of input data about malicious information objects on the Internet and providing their multi-aspect assessment and categorization using machine learning methods. The purpose of the investigation is to improve the efficiency of the detecting process of malicious information on the Internet using the examples of Web-pages classification..*

Keywords: *information objects, malicious information, classifiers, Web-pages, multi-level combination scheme.*

References

1. Hayes P.J., Andersen P.M., Nirenburg I.B., Schmandt L.M. TCS: a shell for content-based text categorization. *Proceedings of the Sixth Conference on Artificial Intelligence Applications, 5–9 May 1990, Santa Barbara, USA*. Piscataway, NJ: IEEE; 1990. vol.1. p.320–326. Available from: https://doi.org/10.1109/CAIA.1990.89206

2. Apté C., Damerau F., Weiss S.M. Automated learning of decision rules for text categorization. *ACM Transactions on Information Systems (TOIS)*. 1994;12(3):233–251. Available from: https://doi.org/10.1145/183422.183423

3. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 1988;24(5):513–523. Available from: https://doi.org/10.1016/0306-4573(88)90021-0

4. Fattah M.A. A Novel Statistical Feature Selection Approach for Text Categorization. *Journal of Information Processing Systems.* 2017;13(5):1397–1409.

5. Lewis D.D., Ringuette M. A Comparison of Two Learning Algorithms for Text Categorization. In: *Third Annual Symposium on Document Analysis and Information Retrieval*. 1994. p.81–93.

6. Joachims T. Text categorization with Support Vector Machines: learning with many relevant features. *Proceedings of the 10th European Conference on Machine Learning, ECML, 21–23 April 1998, Chemnitz, Germany. Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence).* Berlin, Heidelberg: Springer; 1998. vol.1398. p.137–142. Available from: https://doi.org/10.1007/BFb0026683

7. Johnson R., Zhang T. Effective Use of Word Order for Text Categorization with Convolutional Neural Networks. *Proceeding of the Annual Conference of the North American Chapter of the Association for Computational Linguistics " Human Language Technologies", 31 May – 5 June 2015, Denver, USA*. Stroudsburg: Association for Computational Linguistics; 2015. p.103–113. Available from: https://doi.org/10.3115/v1/N15-1011

113. Ghareb A.S., Bakar A.A., Hamdan A.R. Hybrid feature selection based on enhanced genetic algorithm for text categorization. *Expert Systems with Applications*. 2016:49(C):31–47. Available from: https://doi.org/10.1016/j.eswa.2015.12.004

9. Lorena A.C., De Carvalho A.C., Gama J.M.P. A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review*. 2008;30(1-4). Available from: https://doi.org/10.1007/s10462-009-9114-9

10. Kotenko I., Chechulin A., Shorov A., Komashinsky D. Analysis and Evaluation of Web Pages Classification Techniques for Inappropriate Content Blocking. *Proceeding of the 14th Industrial Conference on Data Mining "Advances in Data Mining. Applications and Theoretical Aspects", ICDM, 16–20 July 2014, St. Petersburg, Russia. Lecture Notes in Computer Science.* Cham: Springer; 2014. vol.8557. p.39–54. Available from: https://doi.org/10.1007/978-3-319-08976-8_4

11. Mikolov T., Chen K., Corrado G., Dean J. *Efficient Estimation of Word Representations in Vector Space.* 2013. Available from: https://arxiv.org/pdf/1301.3781 [Accessed 10th April 2019]