Классификация текстовых документов с использованием вероятностной тематической модели*

С.Н. Карпович[I], А.В. Смирнов[II], Н.Н. Тесля[II]

[I] Акционерное общество «Олимп», г. Москва, Россия
[II] Санкт-Петербургский институт информатики и автоматизации РАН, Санкт-Петербург, Россия

Аннотация. Предложен подход к классификации текстовых документов с использованием вероятностной тематической модели, отличающейся тем, что обучающее множество документов представлено экземплярами одного класса. Этот подход позволяет отбирать положительные экземпляры, похожие на заданный класс, из коллекций и потоков текстовых документов. Рассмотрены модели, обучаемые на экземплярах одного класса, решающие задачи классификации в применении к текстовым документам, обозначены их ключевые особенности. Представлена модель классификации Positive Example Based Learning-TM и разработан программный прототип, реализующий классификацию текстовых документов на ее основе. Не имея представления об отрицательных экземплярах документов, она демонстрирует высокую точность классификации, превышающую альтернативные подходы. Экспериментально доказано превосходство Positive Example Based Learning-TM по критерию точности классификации при малом объеме обучающей выборки.

Ключевые слова: классификация, бинарная классификация, тематическое моделирование, обработка текста на естественном языке.

DOI 10.14357/20718594180317

Введение

Классификация текстовых документов является одной из задач машинного обучения. Ее цель заключается в определении группы документов в соответствии со встречаемыми в них словами. Для того чтобы алгоритмы машинного обучения, принимающие на вход числа, могли выполнить классификацию документов, необходимо преобразовать текстовые представления слов в их векторное представление, т.е. провести предварительную векторизацию слов (word embedding). Векторизация слов и документов в настоящий момент является традиционной частью решения задач информационного поиска, тематического моделирования и латентно-семантического анализа [1]. Для обучения классификатора используются специально подготовленные текстовые корпусы, в которых в достаточном количестве представлены документы всех классов. Чтобы отобрать документы одного класса из корпуса текстов, необходимо построить векторное представление документов всего корпуса и решить задачу бинарной классификации.

При решении практических задач возникает потребность в отборе экземпляров одного класса из коллекции или потока текстовых документов. При этом зачастую отсутствуют подходящие корпусы текстов для обучения. В наличии имеется небольшая репрезентативная выборка документов, представляющих искомый класс, который далее будем называть положительным. Текстовые документы, не относящиеся к искомому классу, включаются в

* Работа выполнена при частичной финансовой поддержке РФФИ гранты № 17-07-00327, 17-07-00328.

✉ *Тесля Николай Николаевич e-mail: teslya@iias.spb.su*

другой класс, который называют отрицательным. Векторное пространство, построенное на текстовых документах, представляющих положительный класс, является неполным, так как в этом пространстве не представлены экземпляры отрицательного класса, поэтому существующие алгоритмы классификации текстовых документов не смогут обеспечить приемлемый результат.

В теории машинного обучения существует несколько подходов к решению задачи отбора документов, принадлежащих одному классу [2]. Большинство из них основано на модели опорных векторов, процесс обучения для которых заключается в поиске наименьшей гиперсферы, содержащей большинство элементов обучающей выборки. Применение этих моделей для классификации текстовых документов невозможно, так как, имея неполное векторное представление документов, нет возможности определить положение всех новых слов в этом векторном пространстве.

Цель данной работы заключается в предложении модели классификации с использованием вероятностного тематического моделирования, обучение которой осуществляется с использованием положительных примеров, то есть документов, заведомо относящихся к искомому классу. В работе также описывается создание программного прототипа предложенной модели и его апробация на корпусе текстов SCTM-ru [3].

1. Обзор существующих моделей поиска объектов одного класса и бинарной классификации

Рассматривая упрощенно, задаче поиска объектов одного класса может быть сопоставлена задача определения выбросов или аномалий, иногда называемая одноклассовой классификацией. Модели опорных векторов, используемые для их решения, отделяют основной класс от всех остальных возможных точек, находя гиперсферу с минимальным радиусом вокруг данных из основного класса. Гиперсфера должна содержать максимально возможное количество точек для обучения [4, 5]. Альтернативой является использование гиперплоскости для отделения области с данными от области без данных [6]. В работе [7] предложена робастная модель поиска аномалий, в которой эмпирическое вероятностное распределение заменяется вероятностными распределениями, полученными некоторыми неточными моделями. В работе [8] вводится неточная априорная статистическая информация для улучшения точности классификатора. В работе [9] рассмотрены робастные модели поиска аномалий, в которых метод построения моделей основан на переборе крайних точек множества вероятностей, построенного при помощи «размывания» робастной модели засорения. Ключевым преимуществом данных моделей является более высокая точность по сравнению со стандартными моделями, в случае сравнительно малого размера обучающей выборки. Вышеперечисленные модели поиска аномалий для обучения используют точные, точечные данные [9] и не пригодны для работы с векторным представлением текстовых документов.

Для решения задач поиска текстовых документов одного класса в работе [10] предложена модель Positive Naive Bayes Classifier (PNB), использующая для построения модели классификатора помимо экземпляров, описывающих положительный класс, неразмеченные данные. Ключевым моментом модели PNB является оценка вероятности слова, принадлежащего отрицательному классу, которая оценивается или из отрицательных примеров, или из положительных и немеченых экземпляров классов. Развитие этого подхода предложено в работе [11]. А в работе [12] рассмотрено применение классификации текстовых документов к их потоку, предложен метод ансамблевого динамического классификатора для позитивных и немаркированных данных в потоке текстовых документов. Вышеперечисленные модели используют немаркированные наборы данных для построения модели классификации.

2. Классификация текстовых документов на основе вероятностной тематической модели, построенной на положительных примерах (Positive Example Based Learning – PEBL-TM)

Тематическое моделирование используется для определения тематической принадлежности текстовых документов в коллекциях и потоках этих документов путем построения тематиче-

ской модели коллекции, которая задает отношение между темами и документами в корпусе текстов. Одно из первых упоминаний тематического моделирования появилось в работе [13], в которой предложено вероятностное скрытое семантическое индексирование (PLSI). Наибольшее распространение получила тематическая модель, предложенная в работе [14] – латентное размещение Дирихле (LDA).

Большая часть моделей относится к обучению без учителя, в которых выполняется кластеризация текстовых документов на основе их семантической близости. В работе [15] предложен алгоритм ml-PLSI, на основе которого в настоящей статье будет построена модель классификации текстовых документов с использованием вероятностного тематического моделирования с обучением на положительных примерах.

Модели, разработанные на основе латентного размещения Дирихле, как указано в работе [16], не имеют сильных лингвистических обоснований. При этом классическая модель вероятностного латентно-семантического анализа (PLSA) [13] не связана с какими-либо параметрическими априорными распределениями.

Дадим математическое описание вероятностной тематической модели. Пусть D – коллекция текстовых документов, W - словарь терминов. Каждый документ $d \in D$ представляет собой последовательность терминов n_d и $(w_1,\ldots,w_{n_d})$ из словаря W.

С учетом гипотезы условной независимости $p(w|d,t)=p(w|t)$ по формуле полной вероятности получаем вероятностную модель порождения документа d:

$$p(w|d)=\sum_{t\in T}p(w|d,t)\,p(t|d)=\sum_{t\in T}p(w|t)\,p(t|d)=$$
$$=\sum_{t\in T}\varphi_{wt}\theta_{td}.$$

Для вычисления φ_{wt} и θ_{td} используется ЕМ-алгоритм, являющийся итерационным алгоритмом, в котором на шаге Е вычисляется ожидаемое значение функции правдоподобия, а на шаге М – оценка максимального правдоподобия.

Вероятностная модель появления пары «документ-слово» может быть записана тремя эквивалентными способами:

$$p(d,w)=\sum_{t\in T}p(t)\,p(w|t)\,p(d|t)=$$
$$=\sum_{t\in T}p(d)\,p(w|t)\,p(t|d)=\sum_{t\in T}p(w)\,p(t|w)\,p(d|t),$$

где: T – множество тем, в случае поиска документов — это положительный и отрицательный класс;

$p(t)$ – неизвестное априорное распределение тем в коллекции;

$p(d)$ – априорное распределение на коллекции документов, эмпирическая оценка $p(d)=n_d/n$, где $n=\sum_d n_d$ – суммарная длина всех документов, а n_d – длина документа в словах;

$p(w)$ – априорное распределение на множестве слов, эмпирическая оценка $p(w)=n_w/n$, где n_w – число вхождений слова w во все документы.

Отождествим понятие темы тематической модели и класса в задаче классификации. В качестве данных для обучения классификатора используем экземпляры положительного класса. Априорное распределение на множестве слов $p(w|t)$ демонстрирует вероятность отнесения слова к классу. Чем чаще встречается слово во всех документах обучающего набора, тем оно менее значимо для определения темы. Это свойство тематической модели позволяет точнее настраивать модель классификатора, в сравнении с моделями, учитывающими только наличие слова без учета вероятностной оценки принадлежности к классу.

Для решения задачи поиска документов, принадлежащих определенному классу, используется вероятностная тематическая модель, построенная на экземплярах положительного класса. Текстовый документ, класс которого необходимо определить, разделяется на слова. Для каждого слова определяется вероятность отнесения его к положительному классу. Если слова нет в модели, то вероятность его отнесения к положительному классу будет равна нулю, что не является приемлемым вариантом, поскольку наличие даже одного слова из экземпляров положительного класса в документе, класс которого необходимо определить, будет приводить к присвоению положительного класса всему документу.

Введем понятие штрафа за неопределенность. Считается, что если слова нет в модели, то ему присваивается вероятность отнесения к отрицательному классу. Диапазон значений вероятности–от 0 до 1. Если при решении практической задачи необходима высокая точность в истинно положительных оценках классификатора, то необходимо задать максимально высокий штраф за неопределенность, равный 1. Если важно отобрать как можно больше примеров, пусть и не всегда относящихся к искомому положительному классу, то штраф должен быть меньше 1. Таким образом, значение штрафа подбирается исходя из целей задачи классификации и конкретного набора данных. Алгоритм классификации текстовых документов с использованием вероятностной тематической модели, обученной на положительных примерах, представлен в листинге 1.

Листинг 1: Алгоритм классификации текстовых документов с использованием вероятностной тематической модели, обученной на положительных примерах PEBL-TM.

Вход: коллекция документов D, описывающих положительный класс, штраф за неопределенность $P(negativ)=0$, документы, класс которых необходимо определить d_{new}.

Выход: вероятностная оценка отнесения документа к положительному классу

1. Построить вероятностную тематическую модель на положительных примерах.

2. Для всех $w \in d_{new}$:

- если слова нет в модели, $(w|t)=P(negativ)=1$, $P(positiv)=0$;

- если слово есть в модели $p(w|t)=P(positiv)={n_{dwt}}/{n_{dw}}$,

$P(negativ)=1-P(positiv)$.

3. Считаем вероятность отнесения документа к классу $p(d|t)=\sum_{w\in d}P(positiv)$.

Если обучающая выборка содержит экземпляры отрицательного класса, то они могут быть использованы для построения классификатора на базе вероятностного тематического моделирования. В этом случае задача будет называться бинарной классификацией текстовых документов. При этом новые слова для вероятностной тематической модели должны штрафоваться по аналогии с PEBL-TM. Алгоритм бинарной классификации представлен в листинге 2.

Листинг 2: Алгоритм бинарной классификации текстовых документов с использованием вероятностного тематического моделирования

Вход: коллекция документов D, описывающих положительный класс, включающая экземпляры отрицательного класса, штраф за неопределенность $P(negativ)=0$, документы, класс которых необходимо определить d_{new}.

Выход: вероятностная оценка отнесения документа к положительному классу

1. Построить вероятностную тематическую модель на коллекции документов.

2. Для всех $w \in d_{new}$:

- если слова нет в модели, $p(w|t)=P(negativ)=1$, $P(positiv)=0$.

- если слово есть в модели $p(w|t)=P(positiv)={n_{dwt(positiv)}}/{n_{dw}}$,

$P(negativ)={n_{dwt(negativ)}}/{n_{dw}}$

3. Считаем вероятность отнесения документа к классу $p(d|t)=\sum_{w\in d}P(positiv)$.

Таким образом, если коллекция документов содержит экземпляры двух классов, то следует использовать и отрицательные экземпляры для построения вероятностной тематической модели. Это позволит повысить точность классификатора. Если в коллекции есть только представители положительного класса, то модель PEBL-TM с введенным штрафом, позволит настроить модель классификации.

Следует отметить, что оба предложенных алгоритма могут использоваться как для классификации коллекций текстовых документов, так и для их потоков. Нет необходимости переобучать модель классификатора при анализе потока текстовых документов, хотя такая возможность имеется.

3. Эксперименты с корпусом SCTM-ru

Для разработки прототипа программы, реализующей предложенную модель классификации, был использован язык разработки Python и программные библиотеки для машинного обу-

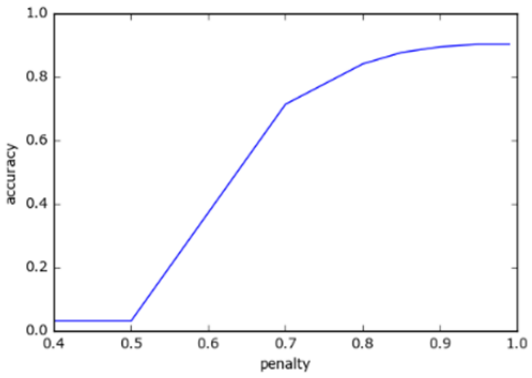

Рис. 1. Изменение метрики accuracy при изменении размера штрафа за неопределенность для модели классификатора с положительными примерами

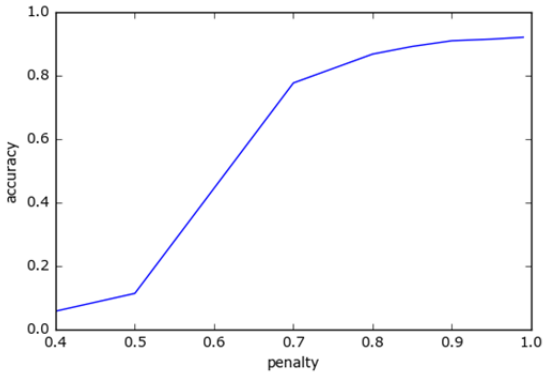

Рис. 2. Изменение метрики accuracy при изменении размера штрафа за неопределенность для модели классификатора с положительными и отрицательными примерами

чения scikit-learn [17] и nltk [18], поставляемые в составе дистрибутива Anaconda. В качестве экспериментальных данных использовался корпус SCTM-ru [3], созданный специально для тестирования задач тематического моделирования. Источником для данного корпуса является международный новостной сайт «Русские Викиновости». Корпус SCTM-ru состоит из 12 тыс. документов, 320 авторов, почти 12000 уникальных категорий. События, описанные в документах, распределены с ноября 2005 года по январь 2017 года. В корпусе SCTM-ru насчитывается 2,5 млн словоупотреблений, состоящих только из русских букв. Словарный состав корпуса составляет 262 тыс. уникальных словоформ.

Каждая новость содержит указанные автором темы. Обычно, перед автором новости не стоит задача перечислить все возможные темы, к которым новость может иметь отношение, тем не менее, указанные темы дают весомые основания полагать, что новость сильно связана с ними.

Для проведения экспериментов были отобраны документы, принадлежащие одной теме, например, «спорт». Весь корпус был разбит на тестовые и обучающие коллекции таким образом, чтобы они не пересекались. Документы из коллекции для обучения, принадлежащие теме «спорт», использовались для обучения классификаторов. Для тестирования классификаторов использовались документы из всех тем тестовой коллекции.

Для оценки качества классификаторов использованы традиционные метрики, а именно: количество истинно положительных ответов, количество ложно положительных ответов, количество истинно отрицательных ответов, количество ложно отрицательных ответов и оценку доли документов, по которым классификатор принял правильное решение к размеру обучающей выборки – точность (accuracy). Лучшим вариантом для решения задачи считается классификатор, показавший лучшую точность.

Для подбора оптимального значения штрафа за неопределенность использовался диапазон значений от 0.4 до 0.99, рассчитывали изменение точности при изменении величины штрафа. Результат для моделей с положительными и с положительными и отрицательными экземплярами представлен на Рис. 1 и Рис. 2. Чем выше штраф, тем выше оценка точности для обеих моделей. С уменьшением штрафа все больше экземпляров классификатор относит к положительному классу, что повышает полноту, но снижает точность классификации.

4. Модель PEBL-TM

Модель классификатора PEBL-TM, была обучена на положительных экземплярах и на экземплярах, включающих примеры из отрицательного класса. Для обучения модели на положительных примерах использовано 117 документов. Для обучения модели на положительных и отрицательных примерах классов использовано 130 документов, из них 117 принадлежат к положительному классу. Метки класса классификатора — это строковые значения, документы, принадлежащие теме спорта, отмеченные меткой «sport», остальные документы отмечены меткой «other». Тестовая коллекция состоит из 1000 заголовков документов,

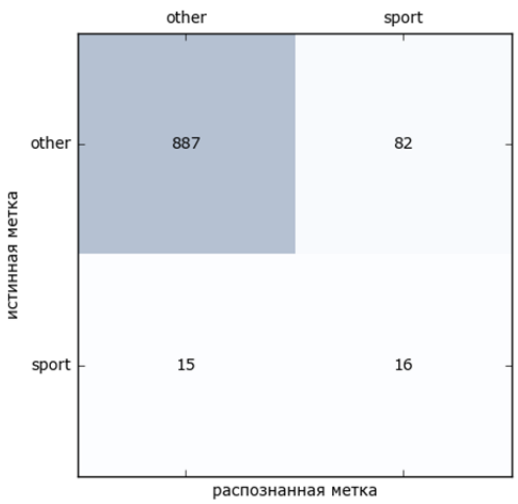

Рис. 3. Результат классификации PEBL-TM только на положительных примерах

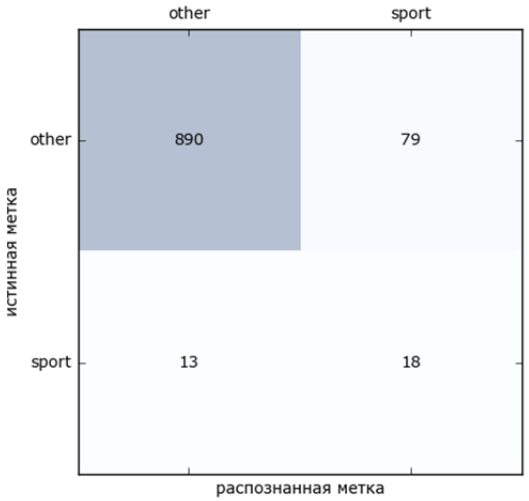

Рис. 4. Результат классификации PEBL-TM на положительных и отрицательных примерах

31 из которых относится к положительному классу. Штраф за неопределенность составляет 0.99 в обеих моделях. Распределение документов тестовой коллекции по классам представлено на Рис. 3 и Рис. 4. Обе модели показали высокую оценку точности (accuracy), равную 0.90. На рисунках видно, что корректно распознаны не все документы из темы спорта. Связано это с репрезентативностью обучающей коллекции. Располагая более репрезентативным набором документов для обучения, можно добиться лучшего качества. Также следует отметить, что часть документов из общей темы классификатор отнес к теме спорта. Это может быть связано, в том числе и с качеством самой коллекции. Темы документов определены их авторами, и для каждого автора не стояла задача отметить все темы, к которым документ может иметь отношение, поэтому вполне возможно, что документы, отнесенные к теме спорта, на самом деле с ней связаны.

5. Модель OneClassSVM

Для оценки использовалась модель одноклассового классификатора SVM (OneClassSVM), входящая в программную библиотеку scikit-learn. Для обучения классификатора положительные экземпляры должны содержать метку «1», отрицательные «-1». Для обучения были использованы 117 документов из обучающей выборки. Для тестирования были использованы 1000 документов, в которых 31 относится к положительному

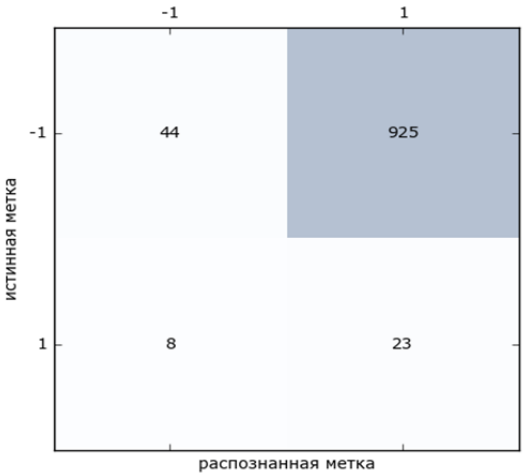

Рис. 5. Результат классификации OneClassSVM

классу. Метрика accuracy равна 0.067, что является значительно более низким результатом по сравнению с моделью PEBL-TM. На Рис. 5 отображено, как классификатор распределил документы тестовой коллекции. Большое количество документов классификатор отнес к теме спорта, что на самом деле не так. Полученные оценки показывают, что этот классификатор не справился с задачей.

6. Модель PositiveNaiveBayesClassifier

Для оценки была использована модель Байесовского одноклассового классификатора (PositiveNaiveBayesClassifier), входящая в программную библиотеку nltk. Для построения

этого классификатора требуются неразмеченные экземпляры (или экземпляры отрицательного класса). Метками класса классификатора являются строковые значения, документы, принадлежащие теме спорта, отмечены как «sport», остальные отмечены меткой «other». Обучающая коллекция содержит 117 заголовков документов, представляющих положительный класс и 100 заголовков, представляющих неразмеченные данные. Тестовая коллекция состоит из 1000 заголовков документов, 31 из которых относится к положительному классу. Точность (accuracy) составляет 0.252, что выше OneClassSVM, но уступает PEBL-TM. На Рис. 6 отображено, как классификатор распределил документы тестовой коллекции. Модель правильно отметила большую часть документов, принадлежащих теме спорта из тестовой коллекции документов, при этом ошибочно отнесла большое количество документов, принадлежащих общей теме, также в тему спорта, что является грубой ошибкой.

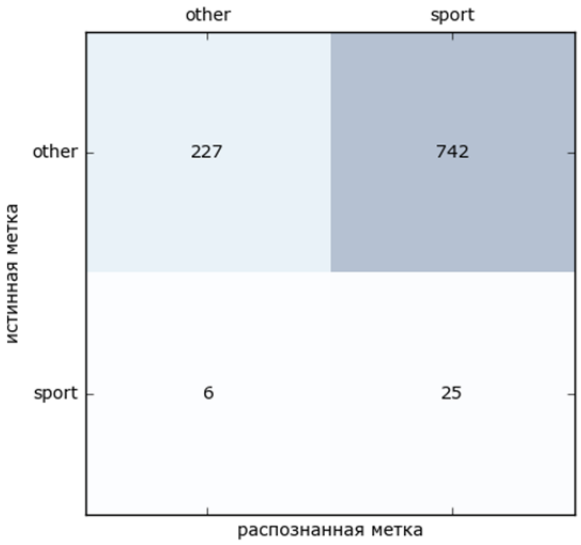

Рис. 6. Результат классификации PositiveNaiveBayesClassifier

Рассмотренные модели классификации, обучаемые на экземплярах одного класса, продемонстрировали низкое качество работы с текстовыми данными. Учитывая, что в обучающей коллекции есть экземпляры отрицательного класса, рассмотрим модели классификации «Случайный лес» (Random forest) и «Логистическая регрессия» (LogisticRegression), успешно зарекомендовавшие себя для работы с текстовыми данными. Для обучения моделей на положительных и отрицательных классах использовано 130 заголовков документов, из них 117 принадлежат к положительному классу. Метки класса классификатора — это числовые значения, документы, принадлежащие теме спорта отмечены «1», остальные документы отмечены меткой «0». Тестовая коллекция состоит из 1000 заголовков документов, 31 из которых относится к положительному классу. Точность (accuracy) у модели Random forest равна 0.031, у модели LogisticRegression – 0.07. Результат классификации представлен на Рис. 7 и Рис. 8. Большинство документов ошибочно отнесено к положительному классу. Причина низкой точности построенных моделей классификации заключается в малом количестве данных для обучения. С увеличением обучающей выборки,

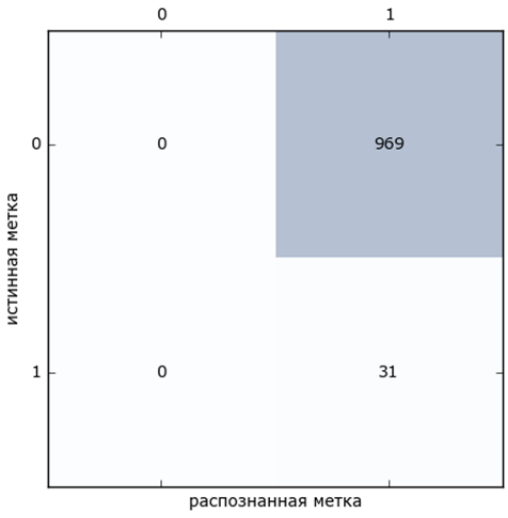

Рис. 7. Результат классификации Random forest

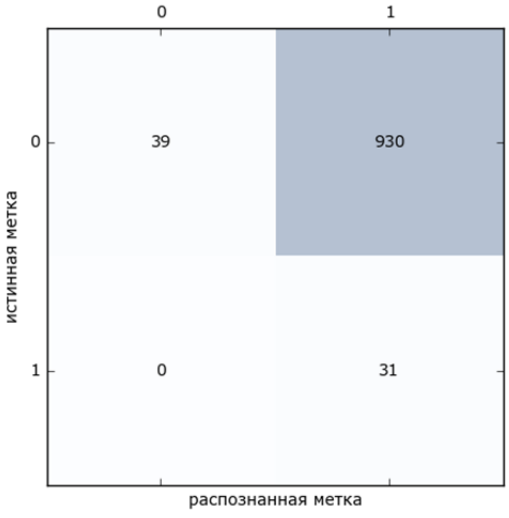

Рис. 8. Результат классификации LogisticRegression

Табл. 1. Сравнение моделей классификации

Модель классификации	**F-мера**	**Accuracy**
PEBL-TM на положительных примерах	**0.24**	**0.90**
PEBL-TM на положительных и отрицательных классах	**0.28**	**0.90**
OneClassSVM	0.04	0.07
PositiveNaiveBayesClassifier	0.06	0.25
Random forest	0.06	0.03
LogisticRegression	0.06	0.07

точность классификации будет расти для всех моделей. Для решения задачи отбора текстовых документов одного класса, с обучением на малой обучающей выборке, такие классификаторы не подходят.

Результаты сравнения моделей классификации представлены в Табл. 1. По полученным данным экспериментов рассчитано значение F-меры – гармонического среднего между точностью (precision) и полнотой (recall) классификации. Значения F-меры и точности (accuracy) моделей PEBL-TM на порядок выше других моделей, участвующих в сравнении.

Заключение

В результате проделанной работы была разработана модель классификации текстовых документов на базе вероятностного тематического моделирования, с использованием только положительных примеров для обучения, а также с использованием положительных и отрицательных примеров. Проведенное сравнение предложенной модели PEBL-TM с существующими моделями поиска аномалий и бинарной классификации демонстрирует перспективность использования вероятностного тематического моделирования в задачах поиска документов, принадлежащих заданному классу.

Вероятностные тематические модели также могут быть использованы в составе ансамблей классификаторов. Для повышения точности PEBL-TM может быть использована регуляризация вероятностной тематической модели.

Результаты проведенных экспериментов и программная реализация модели классификации размещены в свободном доступе по адресу: https://github.com/cimsweb/PEBL-TM/.

Литература

1. Schütze H., Manning C. D., Raghavan P. Introduction to information retrieval. – Cambridge University Press, 2008. – Т. 39, 482 с.
2. Bartkowiak A. M. Anomaly, novelty, one-class classification: a comprehensive introduction //International Journal of Computer Information Systems and Industrial Management Applications. – 2011. – Т. 3. – №. 1. – pp. 61-71.
3. Карпович С. Н. Русскоязычный корпус текстов SCTM-RU для построения тематических моделей //Труды СПИИРАН. – 2015. – Т. 2. – №. 39. – С. 123-142.
4. Tax D., Duin R. Support vector data description. Machine Learning, 2004, no. 54(1), pp. 45–66
5. Tax D., Duin R. Support vector domain description // Pattern Recognition Letters. - 1999. - Vol. 20. - Pp. 1191-1199.
6. Schölkopf B. et al. Estimating the support of a high-dimensional distribution //Neural computation. – 2001. – Т. 13. – №. 7. – С. 1443-1471.
7. Utkin L. A framework for imprecise robust one-class classification models // International Journal of Machine Learning and Cybernetics, 2014, – Т. 5. – №. 3. – С. 379-393. doi: 10.1007/s13042-012-0140-6
8. Utkin L., Zhuk Y. Imprecise prior knowledge incorporating into one-class classification // Knowledge and information systems. – 2014. – Т. 41. – №. 1. – С. 53-76.
9. Уткин Л. В., Жук Ю. А. Робастные модели однокласcовой классификации и крайние точки множества вероятностей //Международная конференция по мягким вычислениям и измерениям. – Федеральное государственное автономное образовательное учреждение высшего образования Санкт-Петербургский государственный электротехнический университет ЛЭТИ им. ВИ Ульянова (Ленина), 2012. – Т. 1. – С. 220-224
10. Denis F., Gilleron R., Tommasi M. Text classification from positive and unlabeled examples //Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02. – 2002. – С. 1927--1934.
11. Denis F. et al. Text classification and co-training from positive and unlabeled examples //Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data. – 2003. – С. 80-87.
12. Pan S., Zhang Y., Li X. Dynamic classifier ensemble for positive unlabeled text stream classification //Knowledge and information systems. – 2012. – Т. 33. – №. 2. – С. 267-287.
13. Hoffman T. Probabilistic Latent Semantic Indexing // Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. — 1999. – С. 50-57.
14. Blei D.M., Ng A.Y., Jordan M. I. Latent Dirichlet Allocation // Journal of Machine Learning Research. — 2003. – Т. 3. – №. Jan. – С. 993-1022.
15. Карпович С. Н. Многозначная классификация текстовых документов с использованием вероятностного тематического моделирования ml-PLSI //Труды СПИИРАН. – 2016. – Т. 4. – №. 47. – С. 92-104.
16. Воронцов К. В., Потапенко А. А. Модификации EM-алгоритма для вероятностного тематического моделирования //Машинное обучение и анализ данных. – 2013. – Т. 1. – №. 6. – С. 657-686.

17. Pedregosa F. et al. Scikit-learn: Machine learning in Python //Journal of machine learning research. – 2011. – Т. 12. – №. Oct. – С. 2825-2830.
18. Bird S., Loper E. NLTK: the natural language toolkit //Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. – Association for Computational Linguistics, 2004. – С. 31.

Text documents classification based on probabilistic topic model

S.N. Karpovich[I], A.V. Smirnov[II], N.N. Teslya[II]

[I] Corporation "Olymp", Moscow, Russia
[II] St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), St. Petersburg, Russia

The paper proposes an approach to the classification of text documents using a probabilistic topic model, with a training set of documents represented by instances of one class. The proposed approach allows selecting positive instances similar to a given class from collections and text document flows. The models learned on instances of one class, solving problems of classification in application to text documents are considered, the key features of such models are indicated. The classification model Positive Example Based Learning-TM is presented and a software prototype is developed, which realizes the classification of text documents based on it. The developed model demonstrates high classification accuracy, which exceeds the alternative approaches. The proposed model as well as existing models was evaluated based on the SCTM-ru text corpora. Experimentally proved the superiority of Positive Example Based Learning-TM by the criterion of classification accuracy with a small size of training set.
Keywords: classification, binary classification, topic model, natural language processing.

DOI 10.14357/20718594180317

References

1. Schütze H., Manning C. D., Raghavan P. Introduction to information retrieval. 2008. 39. 482 p.
2. Bartkowiak A. M. 2011. Anomaly, novelty, one-class classification: a comprehensive introduction. International Journal of Computer Information Systems and Industrial Management Applications. 3(1):61-71.
3. Karpovich S.N. 2015. Russkoyazychnyj korpus tekstov SCTM-RU dlya postroeniya tematicheskikh modelej [The Russian Language Text Corpus for Testing Algorithms of Topic Model]. Trudy SPIIRAN [SPIIRAS Proceedings] 2(39):123-142.
4. Tax D., Duin R. 2004. Support vector data description. Machine Learning. 54(1):45–66
5. Tax D., Duin R. 1999. Support vector domain description. Pattern Recognition Letters. 20:1191-1199.
6. Schölkopf B. et al. 2001. Estimating the support of a high-dimensional distribution. Neural computation. 13(7):1443-1471.
7. Utkin L. 2014. A framework for imprecise robust one-class classification models. International Journal of Machine Learning and Cybernetics. 5(3):379-393.
8. Utkin L., Zhuk Y. 2014. Imprecise prior knowledge incorporating into one-class classification. Knowledge and information systems. 41(1):53-76.
9. Utkin L. V., Zhuk Y A. 2012. Robastnye modeli odnoklassovoj klassifikatsii i krajnie tochki mnozhestva veroyatnostej [Robust models of the one-class classification and extreme points of the probability set]. Mezhdunarodnaya konferentsiya po myagkim vychisleniyam i izmereniyam [International Conference on Soft Computing and Measurement] 1:220-224.
10. Denis F., Gilleron R., Tommasi M. 2002. Text classification from positive and unlabeled examples. Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, IPMU'02. 1927-1934.
11. Denis F. et al. 2003. Text classification and co-training from positive and unlabeled examples. Proceedings of the ICML 2003 workshop: the continuum from labeled to unlabeled data. 80-87.
12. Pan S., Zhang Y., Li X. 2012. Dynamic classifier ensemble for positive unlabeled text stream classification. Knowledge and information systems. 33(2):267-287.
13. Hoffman T. 1999. Probabilistic Latent Semantic Indexing. Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval. 50-57.
14. Blei D.M., Ng A.Y., Jordan M. I. 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research. 3:993-1022.
15. Karpovich S.N. 2016. Mnogoznachnaya klassifikatsiya tekstovykh dokumentov s ispol'zovaniem veroyatnostnogo tematicheskogo modelirovaniya ml-PLSI [Multi-label classification of text documents using probabilistic topic modeling]. Trudy SPIIRAN [SPIIRAS Proceedings] 4(47):92-104.
16. Vorontsov K. V., Potapenko A. A. 2013 Modifikatsii EM-algoritma dlya veroyatnostnogo tematicheskogo modelirovaniya [EM-like algorithms for probabilistic topic modeling]. Mashinnoe obuchenie i analiz dannykh [Machine Learning and Data Analysis]. 1(6):657-686.
17. Pedregosa F. et al. 2011. Scikit-learn: Machine learning in Python. Journal of machine learning research. 12:2825-2830.
18. Bird S., Loper E. 2004 NLTK: the natural language toolkit //Proceedings of the ACL 2004 on Interactive poster and demonstration sessions. – Association for Computational Linguistics. 31.