УДК 004.93'11

КЛАССИФИКАЦИЯ ТЕКСТОВ НА ОСНОВЕ ПРИБЛИЖЕННЫХ ОЦЕНОК ВЕРОЯТНОСТЕЙ КЛАССОВ

А. С. Солодухин

Омский государственный технический университет

В данной работе предлагается метод классификации текстов на основе приближенных оценок условных распределений вероятностей классов. Суть метода заключается в представлении наборов признаков и класса текста как совокупность одновременных событий и приближенной оценки вероятностных зависимостей между признаками и классами текстов.

ВВЕДЕНИЕ

Автоматическая классификация (категоризация) текстов в предопределенные категории получила большое внимание в последние 10—15 лет, из-за увеличения имеющихся документов в цифровом виде и возникшей потребности хранить их в организованном виде [1]. Доминирующий подход к решению этой задачи у исследователей базируется на технологиях машинного обучения: автоматическое построение классификатора путем обучения на основе обучающего множества предварительно классифицированных документов, характеризующих классы (категории). Преимущество данного подхода перед подходом инженерии знаний (заключающимся в ручном определении классификатора экспертом) — высокая эффективность, значительное сокращение труда эксперта, применение в различных областях.

Очевидно, что невозможно создать метод, который позволит классифицировать документы с абсолютной точностью, поскольку классификация чаще всего бывает субъективной. По этой причине всегда будет существовать научное направление совершенствования методов классификации и построения классификаторов.

Известные классификаторы, имея достаточное количество необходимых примеров в обучающей выборке, зачастую не обеспечивают требуемую точность классификации [1, 2].

Известно, что практическими приложениями методов категоризации текстов являются:

— фильтрация документов, распознавание спама;

— автоматическое аннотирование;

— снятие неоднозначности (автоматические переводчики);

— составление интернет-каталогов;

— классификация новостей;

— распределение рекламы;

— персональные новости.

В имеющихся публикациях, посвященных задаче классификации текстов, рассматриваются различные методы. Среди них вероятностные методы [3, 4], деревья решений [5], правила решений [6], регрессионные методы [1], поиск k ближайших соседей [1, 2], искусственные нейронные сети [7, 8], метод опорных векторов [9, 10], энтропийный метод [11, 12].

Ниже используются следующие определения:

— текстовый документ (текст) — электронный файл, содержащий только текст без служебной информации;

— класс (категория) — параметр текста который характеризует содержащийся в тексте смысл;

— классификация (категоризация) текстов — установка соответствия класса текстовому документу;

— классификатор — алгоритм, ставящий в соответствие класс текстовому документу.

В содержательной постановке задачи классификации текстов дано множество предопределенных произвольных категорий текстовых документов и множество предварительно классифицированных документов-примеров, необходимо на основе имеющегося множества предварительно классифицированных документов-примеров построить некоторый классификатор, который для поданного на вход текстового документа определял его класс с некоторой степенью точности. Причем классификация должна выполняться только на

основе анализа, содержащегося текста в документе.

Таким образом, задачей категоризации текстов [1,2] является определение булевого значения для каждой пары $(d_j, c_i) \in D \times C$, где D — множество документов и $C = \{c_0, c_1, \ldots, c_{|C|}\}$ — множество предопределенных категорий (классов). Значение «Истина» (T) говорит о том, что документ d_j соответствует категории c_i, значение «Ложь» (F) — не соответствует. Более формально задачей является нахождение приближения неизвестной целевой функции $\Phi : D \times C \to \{T, F\}$, которая описывает, как классифицируются документы, и называется классификатором (правилом решения, гипотезами, моделью). Причем:

— категории представляют собой только символические метки, и нет никаких дополнительных знаний процедурного или декларативного характера;

— отсутствуют какие-либо, дополнительные знания для целей классификации, классификация должна быть выполнена только на основании содержания документов.

Поиск функции Ф осуществляется с помощью некоторого алгоритма обучения на основе предварительно классифицированных документов D (обучающее множество примеров, тренировочный набор):

$$\{(d_1, c_1), (d_2, c_2), \ldots, (d_n, c_n)\}, d_i \in D, c_i \in C.$$

Для решения задачи обучения и классификации документ d_j, при помощи некоторой процедуры индексации [13, 14], суть которой заключается в поиске признаков в документе, представляется в виде вектора признаков

$$x_j = (w_{1j}, \ldots, w_{|T|j}),$$

где w_{1j} булевое или весовое значение, характеризующее некоторый признак в документе. Обычно признаком является терм — подстрока, выбранная при помощи некоторого алгоритма [там же], суть которого заключается в выборе множества термов, обеспечивающего достаточно эффективную классификацию, на основе обучающего множества документов.

1. ПРЕДЛАГАЕМЫЙ МЕТОД КЛАССИФИКАЦИИ ТЕКСТОВ

Суть предлагаемого метода заключается в вычислении приближенной оценки распределений вероятностей классов. Пусть имеется множество бинарных признаков w_k и множество классов c_i, соответствующие множеству текстовых документов d_j. Рассмотрим эти множества как зависимые совместные случайные события. Текстовый документ d_j, имеющий вектор признаков x_j, класс которого известен c_i, может рассматриваться как совокупность одновременных событий (x_j, c_i). Пусть (x_j, c_i) является статистическим экспериментом тогда можно рассчитать статистические оценки условных вероятностей классов c_i, условиями которых являются x_j.

Рассмотрим простейший случай классификации, пусть имеются совместные зависимые случайные события: класс a и некоторые признаки b и c, необходимо вычислить вероятность класса по данным признакам $P(a / bc)$. Допустим, имеется точная информация обо всех возможных исходах, которая позволяет вычислить точно значение вероятности $P(a / bc)$. Например, множества возможных исходов могут выглядеть так, как представлено на рис. 1.

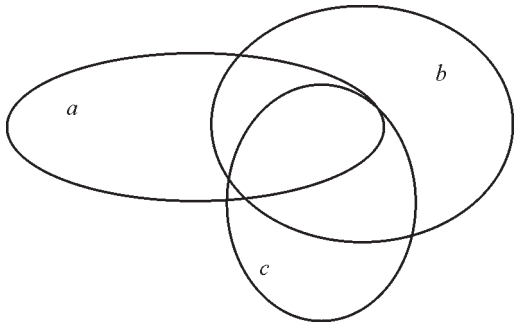

Рис. 1. Множества возможных исходов при классификации объектов

В случае, представленном на рис. 2 можно быть уверенным в том, что $P(a / bc) = 0$ — события a, b и c не могут возникнуть одновременно, говорит о том, что к классу a не относится объект, обладающий признаками b и c одновременно.

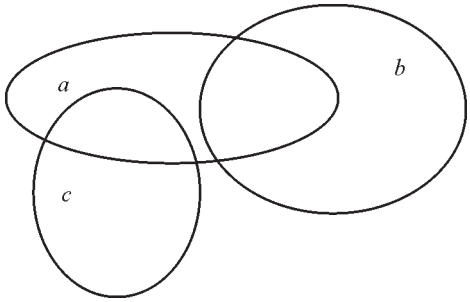

Рис. 2. Множества возможных исходов при классификации объектов с $P(a / bc) = 0$

Рассмотрим ситуацию, когда вероятность оценивается экспериментальным путем. В результате проведения серии испытаний были оценены множества всех возможных исходов, которые изображены на рис. 3, где сплошной линией указаны точные возможные множества исходов, а пунктирной линией оцененные в данный момент. Тогда при вычислении оценки условной вероятности

$$P^*(a/bc) = \frac{P(abc)}{P(bc)} = \frac{C(abc)}{C(bc)}$$

может случиться, что $P^*(a\ /\ bc) = 0$, где $C(abc)$ и $C(bc)$ — количество встреч событий (abc) и (bc), соответственно, по причине $C(abc) = 0$. Проблема «молодой статистики» не позволяет всегда вычислять оценку $P^*(a\ /\ bc)$ приемлемую для практического применения. Для вычисления оценки $P^*(a\ /\ bc)$ в условиях ограниченного количества опытов необходимо разработать приближенные методы. Суть проблемы «молодой статистики» заключается в том, что в текущий момент времени проведенного количества опытов недостаточно для вычисления оценки $P^*(a\ /\ bc)$, очевидно, что при дальнейших опытах оценка $P^*(a\ /\ bc)$ будет приближаться к точному значению $P(a\ /\ bc)$. Возникает задача прогноза значения $P^*(a\ /\ bc)$, не зависимо от того, закончена ли серия опытов или нет.

На основе проведенного анализа работ [15—23] можно сделать вывод, что на данный момент не существует достаточно теоретически обоснованного подхода к эффективному приближенному вычислению оценки $P^*(a\ /\ bc)$ при молодой статистике.

Можно предположить, что $P^*(a\ /\ bc)$ или $C(abc)$ каким-либо образом, приближенно, зависит от $C(a)$, $C(b)$, $C(c)$, $C(ab)$, $C(bc)$, $C(ac)$ и может быть на основании их приближенно вычислена. Но справедливы только неравенства:

$$C(a) \geq C(ab);\ \ C(c) \geq C(bc);\ \ C(ab) \geq C(abc);$$
$$C(b) \geq C(ab);\ \ C(a) \geq C(ac);\ \ C(bc) \geq C(abc);$$
$$C(b) \geq C(bc);\ \ C(c) \geq C(ac);\ \ C(ac) \geq C(abc).$$

Это говорит о том, что вероятность $P^*(a\ /\ bc)$ при любых $C(a)$, $C(b)$, $C(c)$, $C(ab)$, $C(bc)$, $C(ac)$ может быть в интервале от 0 до 1, т.е. любой.

Можно предположить, что прогноз значения $P^*(a\ /\ bc)$ в случае, когда отсутствует достаточная статистика $C(abc)$, должен основываться на некоторой модели воздействия на событие-следствие событий-условий и взаимодействия событий-условий. Имеются следующие вопросы:

— как события-условия воздействуют на событие-следствие (как это выразить математически);

— как события-условия взаимодействуют между собой (как это выразить математически).

Очевидно что количество встреч событий $C(c_i\ x_j)$ на обучающей выборке тестовых документов будет чаще близко к 1 поскольку при правильно выбранных признаках w_k, векторы признаков x_j должны быть более менее уникальными для каждого документа d_j. По той же самой причине оценки $P^*(c_i/x_j)$ для текстовых документов, которые необходимо классифицировать чаще будут равны 0, поскольку $C(c_i\ x_j)$ скорее всего будет равно 0. Для решения данной проблемы необходимо разработать приближенный метод вычисления $P^*(c_i/x_j)$.

Таким образом, поставлена задача статистической оценки распределений вероятностей классов текстовых документов, рассмотрим приближенный метод ее решения.

На основе обучающей выборки вычислим множество оценок условных вероятностей $P(c_i/f)$, где $f \in T$ — любое встречаемое в документах d_j подмножество событий w_k. Поскольку известно, что события c_i несовместные, то можно сказать, что результатом являются условные распределения вероятностей $P(c_i/f)$, где $i = (1,\ldots,|C|)$ и $f \in T$. Для классификации необходимо получить распределение вероятностей $P(c_i/K)$, где $i = (1,\ldots,|C|)$ и K некоторое условие, характеризующее всю накопленную статистику в процессе обучения. Другими словами, проблема заключается в следующем: как получить единственное классифицирующее распределе-

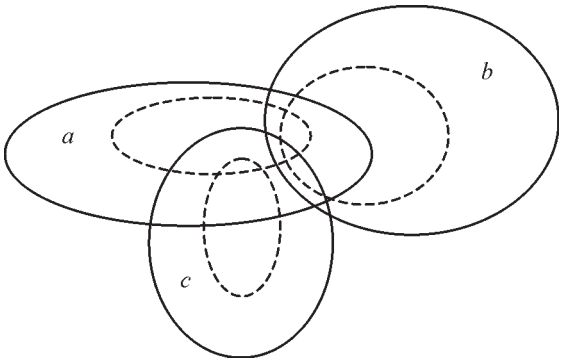

Рис. 3. Множества возможных исходов при классификации объектов и исходы, полученные опытным путем (отмечено пунктирной линией)

ние вероятностей из множества имеющихся условных распределений вероятностей (с условиями $f \in T$).

Данная проблема встречается в методах сжатия информации, основанных на контекстном моделировании [25—32]. Данные методы используют энтропийное кодирование, в котором для кодирования следующего символа необходимо иметь распределение вероятностей всех возможных в данный момент символов. Контекстное моделирование основано на поиске условных вероятностей символов при условии предшествующего контекста. Данные условные вероятности используются для формирования распределения вероятностей прогнозируемого следующего символа. Если назвать прогнозируемый символ классом, а контекстные условия признаками, то данные задачи имеют много общего. Следовательно, подходы к решению данной проблемы в методах сжатия можно использовать в рассматриваемом методе классификации.

Выберем модель решения — смешивание условных распределений вероятностей. Предположим, что каждая оценка $P(c_i/f)$ имеет приближенное к $P^*(c_i/x_j)$ в некоторой степени значение. Пусть каждая оценка $P(c_i/f)$ имеет свой вес, а оценка $P^*(c_i/x_j)$ вычисляется как сумма оценок $P(c_i/f)$, умноженных на свои веса.

На основе выбранной модели рассмотрим предлагаемый метод решения задачи классификации текстов. Одним из методов смешивания распределений вероятностей в методах сжатия является сложение всех распределений вероятностей, умноженных на коэффициент *weight* (вес распределения), и последующая их нормализация (рис. 4).

Где $S = \sum_{i=1}^{|C|} S_p(c_i)$ — нормирующий коэффициент.

Главной проблемой смешивания является определение весов $weight_k$. Одним из вариантов может быть смешивание с весами $weight_k$=1. Возможен вариант смешивания, когда веса $weight_k$ находятся каким-либо методом оптимизации.

Пусть имеется множество весов $W=\{weight_k\}$ и множество обучающих документов $D=\{d_j\}$. Для заданных весов W можно определить количество правильно классифицированных документов S из D путем непосредственной классификации каждого документа d_j, используя

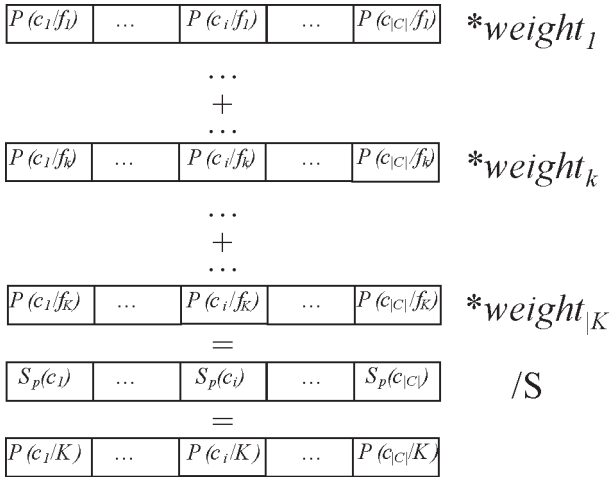

Рис. 4. Модель приближенного вычисления оценок распределения вероятностей

классификатор. Необходимо найти такие W, чтобы S было максимальным. Данную задачу можно решить при помощи метода градиентного спуска.

Поскольку количество весов $weight_k$ обычно примерно порядка 10^4 (количество возможных слов в классифицируемых текстах), то поиск их значений методами оптимизации достаточно ресурсоемкий. Если задача может эффективно решаться без постоянного дообучения классификатора на пополняемой обучаемой выборке, то приведенный метод поиска весов может оказаться приемлемым. С целью экономии процессорного времени возможен вариант разделения весов $weight_k$ на группы с последующей аналогичной оптимизацией групп весов.

В итоге сделанное приближение позволило разработать метод классификации текстов на основе вычисления приближенной статистической оценки распределения вероятностей классов заданного документа. Данный подход имеет достаточно общие теоретические основания, возможно, поэтому он может применяться для решения других задач, которые могут быть сведены к вычислению оценки условной вероятности, вычисляемой обычными методами с большой погрешностью.

2. ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Для проведения экспериментов были разработаны алгоритмы и необходимое программное обеспечение. В исследованиях использовалось множество текстовых документов «Ling-Spam corpus» (составитель Ion Androutsopoulos). Данное множество документов является стан-

дартным при исследовании алгоритмов классификации текстов. Результаты исследований других алгоритмов классификации текстов на данном множестве текстовых документов можно найти во многих публикациях [1, 2]. «Ling-Spam corpus» представляет собой множество текстовых документов, электронных писем, разделенных на два класса: легальная почта (Legal — 2112 писем), нежелательная почта (Spam — 781 письмо). Для оценки эффективности классификатора используем меры, описанные в литературе [1, 2]. Данные меры оценки широко применяются в публикациях, поэтому по ним можно проводить сравнения с аналогичными методами классификации текстов.

Для проведения эксперимента из текстовой выборки были выбраны документы, содержащиеся в папке «bare». В качестве обучающего множества были выбраны все подпапки кроме «part10», соответственно «part10» использовалась для контрольной классификации. Результаты экспериментов представлены в табл.

Таблица

Результаты экспериментов с использованием выборки «Ling-Spam corpus»

Категория	precision	recall
Spam	0,959	0,959
Legal	0,991	0,991

Значения мер оценок эффективности классификации:

— микро-усредненный *recall* 0,986;

— микро-усредненный *precision* 0,986;

— сбалансированная точка системы 0,986.

При другом разделении «Ling-Spam corpus» на обучающее множество и множество для тестирования качества классификации значения мер оценок отличаются не более чем на 5 %.

Разработанный метод обеспечивает более высокое качество классификации, чем качество известных эффективных методов (SvmLight [Joachims 1998], LLSF [Yang 1999], NNET [Yang and Liu 1999], k-NN [Yang 1999]) на данной текстовой выборке.

РЕЗУЛЬТАТЫ

Предложен подход классификации текстовых документов в предопределенные классы с использованием статистической оценки распределений вероятностей классов. Разработан новый метод классификации текстов с использованием метода приближенного вычисления распределений оценок вероятностей, ранее не применявшегося в области классификации текстов. Проведены экспериментальные исследования разработанного метода и установлено, что полученный в работе метод позволяет повысить качество классификации текстов.

СПИСОК ЛИТЕРАТУРЫ

1. *Fabrizio Sebastiani.* Machine Learning in Automated Text Categorization. ACM Computing Surveys, Vol. 34, No. 1, March 2002, P. 1—47.

2. *Kjersti Aas and Line Eikvil.* Text Categorisation: A Survey.

3. *Lewis, D. D.* 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, Germany, 1998), 4—15.

4. *Susana Eyheramendy, David D. Lewis, David Madigan.* On the Naïve Bayes Model for Text Categorisation.

5. *Mitchell, T.M.* 1996. Machine Learning. McGraw Hill, New York, NY.

6. C. Apte, F. Damerau, and S.M. Weiss. Automated Learning of Decision Rules for Text Categorization. ACM Transactions on Information Systems, 1994.

7. *Lam, S. L. and Lee, D. L.* 1999. Feature reduction for neural network based text categorization.In Proceedings of DASFAA-99, 6th IEEE International Conference on Database Advanced Systems for Advanced Application (Hsinchu, Taiwan, 1999), 195—202.

8. *Yang, Y. and Liu, X.* 1999. A re-examination of text categorization methods. In Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval (Berkeley, CA, 1999), 42—49.

9. *Joachims,* T. 1998. Text categorization with support vector machines: learning with many relevant features. In Proceedings of ECML-98, 10th European Conference on Machine Learning (Chemnitz, Germany, 1998), 137—142.

10. *Joachims, T.* 1999. Transductive inference for text classification using support vector machines. In Proceedings of ICML-99, 16th International Conference on Machine.

11. *Frank E., Chui C., Witten I.H.* Text categorization using compression models. // Proc Data Compression Conference, edited by J.A. Storer, et al., Snowbird, Utah, March. IEEE Press, Los Alamitos, pp. 555.

12. *Хмелев Д.В.* Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение [Электронный ресурс] / Д. В. Хмелев Режим доступа: http://compression.graphicon.ru/download/articles/classif/intro.html, свободный.

13. *David D. Lewis*. Feature Selection and Feature Extraction for Text Categorisation.

14. *Sam Scott, Stan Matwin*. Feature Engineering for Text Classification.

15. *Бернштейн С.Н.* Теория вероятностей / С. Н. Бернштейн. — М. : Гос. изд-во, 1927. — 367 с.

16. *Лоэв М.* Теория вероятностей / М. Лоэв. — М. : Изд-во иностранной литературы, 1962. — 721 с.

17. *Бернулли Я.* О законе больших чисел / Я. Бернулли. — М.: Наука. Гл. ред. физ.-мат. лит., 1986. — 176 с.

18. *Колмогоров А.Н.* Введение в теорию вероятностей / А. Н. Колмогоров, И. Г. Журбенко, А. В. Прохоров — М.: Наука. Гл. ред. физ.-мат. лит., 1982. — 160 с. — (Библиотечка «Квант». Вып. 23).

19. *Гмурман В.Е.* Теория вероятностей и математическая статистика : учеб. пособие для вузов / В. Е. Гмурман. — М. : Высш. шк., 2003. — 479 с.

20. *Пуанкаре А.* Теория вероятностей / А. Пуанкаре — Ижевск : Ижевская республиканская типография, 1999. — 280 с.

21. *Савельев Л.Я.* Элементарная теория вероятностей : учеб. пособие. / Л. Я. Савельев : Новосибирск, Новосибирский государственный университет. 2005. — 190 с.

22. *Колмогоров А.Н.* Основные понятия теории вероятностей / А. Н. Колмогоров — М. : Главная редакция общетехнической литературы и номографии, 1936. — 82 с.

23. *Кибзун А.И.* Теория вероятностей и математическая статистика. Базовый курс с примерами и задачами / А. И. Кибзун, Е. Р. Горяинова, А. В. Наумов, А. Н. Сиротин — М. : Физматлит, 2002. — 224 с.

24. *Cleary J.G. and Witten I.H.* Data compression using adaptive coding and partial string matching. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/cleary_witten_1984_pdf.rar, свободный.

25. *Bell T., Witten I.H., Cleary J.G.* Modeling for text compression. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/rev_univ/bell_1989_modeling_pdf.rar, свободный.

26. *Bloom C.* Solving the problems of context modeling. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/bloom_1996_ppmz_pdf.rar, свободный.

27. *Teahan W.J.* Probability estimation for PPM. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/teahan_1995_probest_pdf.rar, свободный.

28. *Howard P.G.* The design and analysis of efficient lossless data compression systems. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/howard_phd_1993_pdf.rar, свободный.

29. *Bunton S.* On-line stochastic processes in data compression. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/bunton_phd_1996_pdf.rar, свободный.

30. *Witten I.H. and Bell T.C.* The zero-frequency problem: estimating the probabilities of novel events in adaptive text compression. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/witten_bell_1991_zerofreq_pdf.rar, свободный.

31. *Willems F., Shtarkov Y., Tjalkens T.* The context tree weighting method: basic properties. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/cm/willems_1995_ctw_basic_pdf.rar, свободный.

32. *Cleary J.G., Teahan W.J., Witten I.H.* Unbounded length contexts for PPM. [Электронный ресурс]. Режим доступа: http://compression.graphicon.ru/download/articles/ppm/cleary_teahan_1997cj_pdf.rar, свободный.