ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2014 Управление, вычислительная техника и информатика № 4 (29)

ОБРАБОТКА ИНФОРМАЦИИ

УДК 519.237.8: 81’322.2

М.Ф. Ашуров

СРАВНЕНИЕ ПОТОКОВЫХ МЕТОДОВ КЛАССИФИКАЦИИ ТЕКСТОВ ХУДОЖЕСТВЕННОЙ ЛИТЕРАТУРЫ НА ОСНОВЕ СЖАТИЯ ИНФОРМАЦИИ И ПОДСЧЕТА ПОДСТРОК

Рассматривается задача сравнения качества методов классификации текстов на основе *R*-меры и PPM метода. Оценка качества классификаторов для каждой тестовой выборки проводилась с использованием *F*-меры. Проведена классификация по авторам на двух выбранных тестовых выборках с характерными особенностями, связанными с периодами написания художественных произведений. Для каждого классификатора выявлены факторы, которые влияют на снижение качества классификации.
Ключевые слова: классификация текстов; художественные произведения; *R*-мера; PPM метод; *F*-мера.

В современном мире возрастающий объём информации в электронном виде всё больше нуждается в классификации для лучшего хранения и дальнейшей обработки. Ручная классификация при таком объеме текстов будет слишком затратной по времени и человеческим усилиям. Данную проблему призвана решить компьютерная автоматизированная классификация, на основе которой компьютерные комплексы могут справляться с большими объемами информации.

1. Потоковые методы классификации

Многие авторы [1–4] среди большого количества разработанных на сегодняшний день методов решения этой задачи разделяют их на признаковые и потоковые. Признаковые методы (feature-based approaches) основаны на использовании численного представления текстов – векторами значений выбранных признаков. Основной недостаток применения признаковых методов заключён в сложности и трудоемкости определения необходимого и достаточного набора признаков, по которому будет проходить классификация.

В отличие от признаковых методов потоковые методы (stream-based approaches) не требуют каких-либо преобразований и изменений в тексте. Они непосредственно используют элементы текста, т.е. текст X рассматривается как последовательность (поток) из n элементов $x_1, x_2, \ldots x_n$ некоторого алфавита Q, при этом длина текстовой строки $n = |X|$. В качестве элемента текста x_n может быть выбран одиночный текстовый символ, слово, грамматический класс, любая группировка символов текста.

Среди потоковых методов О.Г. Шевелев [5] выделяет два основных направления:

– подсчет повторений строк (*R*-, *C*- и другие меры);

– сжатие информации (off-the-shelf, PPM).

Одной из отличительных особенностей первого направления является то, что результаты обработки могут быть представлены в естественном для человека виде и проанализированы позднее. Стоит отметить, что Д.В. Хмелёв в работе [6] приводит результаты сравнения для *R*-меры и метода PPM, однако данное сравнение происходит на текстах новостной ленты – публицистических статьях, в которых авторский стиль проявляется весьма ярко. Для более полной оценки этих направлений требуется провести их сравнение на текстах разного типа.

2. Усечённая *R*-мера

Методы, предложенные Хмелёвым и Тианом [2], строятся по изложенным выше принципам, но их отличительной особенностью в получении самой меры является результат подсчета определенных подстрок исследуемого текста, которые есть в супертексте (конкатенации всех текстов класса).

Усечённая *R*-мера близости [1] учитывает все возможные повторения всех подстрок длин от k_1 до k_2 испытуемого текста длины n в супертексте (в отличие от неусечённой, для которой $k_1 = 1$, $k_2 = n$):

$$R(X \mid S) = r(X \mid S) / N,$$

$$r\left(X \mid S\right) = \sum_{k=k_1}^{k_2} c_k\left(X \mid S\right),$$

$$N = \left(2(n+1) - (k_1 + k_2)\right)\left(k_2 - k_1 + 1\right) / 2,$$

$$c_k\left(X \mid S\right) = \sum_{i=k}^{n} c\left(x_{i-k+1} \ldots x_n \mid S\right),$$

$$c\left(x_{i-k+1} \ldots x_n \mid S\right) = \begin{cases} 1, x_{i-k+1} \ldots x_n \subset S, \\ 0, x_{i-k+1} \ldots x_n \not\subset S, \end{cases}$$

где X – испытуемый текст; S – супертекст исследуемого класса; k – длина подстроки поиска; N – число подстрок.

Для ускорения вычисления *R*-меры можно использовать суффиксные массивы или суффиксные деревья.

3. PPM метод

Изначально PPM (prediction by partial matching) метод – метод контекстно-зависимого моделирования ограниченного порядка (finite-context modeling), позволяющий оценить вероятность символа в зависимости от предыдущих символов. Префикс строки, непосредственно предшествующий текущему символу, называется контекстом. Если для оценки вероятности используется контекст длины *N*, то данный случай является контекстно-ограниченной моделью степени *N* или порядка *N* (order-*N*, O-*N*). Чтобы получить хорошую оценку вероятности символа, необходимо учитывать контексты разных длин, т.е. PPM может быть представлена как вариант стратегии перемешивания: оценки вероятностей, сделанные на основании контекстов разных длин, объединяются в одну общую вероятность. Полученные оценки кодируются любым энтропийным кодером, например любым арифметическим кодером, за счет чего и происходит сжатие текста.

Алгоритм классификации первоначально требует получить сжатый PPM методом супертекст класса, после чего происходит конкатенация супертекста и испытуемого текста и дальнейшее их сжатие тем же способом. В простейшем наивном подходе мера близости исследуемого текста к классу вычисляется как разница длин сжатой конкатенации супертекста и испытуемого текста относительно сжатого супертекста. Выбор принадлежности текста к классу строится на выборе наименьшей разницы длин вышеописанного способа, т.е. используется идея наименьшей энтропии между текстами класса и испытуемого текста.

4. Постановка задачи

В 2004 г. в Томском государственном университете (ТГУ) научным коллективом под руководством В.В. Поддубного был основан проект программного комплекса «СтилеАнализатор» [7], используемый для выполнения, прежде всего, стилевого анализа размеченных корпусов текстов. Первоначально в «СтилеАнализаторе» был реализован анализ текстов с помощью различных признаковых методов, но затем было принято решение реализовать в комплексе также потоковые методы и провести сравнительный анализ этих методов.

В данной работе анализ обоих классификаторов, основанных на усечённой *R*-мере и сжатия PPM, проводился нами на двух тестовых выборках. Первая выборка состоит из текстов авторов русской художественной прозы XIX в. Всего было использовано 9 авторов, представлявших авторские классы, около 100 произведений которых были выбраны в качестве «обучающей» выборки. Вторая выборка строится из текстов авторов русской художественной прозы 90-х гг. XX в. Использовалось 112 текстов 21 автора. Образованные этими произведениями супертексты каждого класса были уравнены по объему. Тексты тестовой (контрольной) выборки каждого класса были составлены тоже равного объема, порядка 100 тыс. символов каждый, и в равном количестве, при этом 2 из них были получены из текстов обучающей выборки, а оставшаяся часть – из других произведений, не участвовавших в обучении.

Для проведения тестирования качества потоковых методов классификации был спроектирован и реализован на языке C# программный модуль, позволяющий проводить классификацию текстов на основе *R*-меры и сжатия PPM. Проблема выбора параметров k_1 и k_2 для усеченной *R*-меры решалась с учётом особенностей текстов на естественных языках. Минимальная длина подстроки k_1 равна 10 символам, так как более короткие подстроки начинают совпадать с большинством слов русской речи, обычно используемых всеми авторами, что не позволяет выделить стилевые особенности разных авторов. Максимальная длина подстроки k_2 равна 45 символам – такая длина может включать в себе даже словосочетание из 3 или 4 слов. Использование большей длины подстроки видится сомнительным, так как повтор такой длинной фразы у разных авторов представляется крайне маловероятным (проблема плагиата полностью исключалась из рассмотрения). Модуль классификатора на основе PPM метода использует алгоритм PPMD порядка 5, реализованный в RAR-архиваторе.

5. Оценка качества классификатора

Качество классификации по каждому классу оценивалось по текстам контрольной выборки *F*-мерой – средним гармоническим между полнотой (Recall) *r* (долей текстов, правильно приписываемых классу из всех текстов этого класса) и точностью (Precision) *p* (долей текстов, правильно приписываемых классу из всех текстов, приписываемых этому классу) [1]:

$$F = 2\frac{p \times r}{p + r}.$$

Среднее значение *F*-меры по всем классам принималось за оценку качества потоковой классификации каждого классификатора в целом.

6. Результаты сравнения

Относительно первой тестовой выборки, состоящей из 9 классов, для усеченной *R*-меры были получены следующие характеристики, приведенные в табл. 1.

Таблица 1

Точность, полнота и *F*-мера на текстах авторов XIX в. для *R*-меры

Автор	Precision	Recall	F-measure
Chehov Anton	0,60	0,75	0,67
Dostoevskii Fedor	1,00	0,42	0,59
Gogol Nikolay	0,92	1,00	0,96
Goncharov Ivan	0,67	1,00	0,80
Kuprin Alexandr	1,00	0,50	0,67
Leskov Nikolai	0,71	0,83	0,77
Saltikov-Shedrin Mihail	0,40	0,17	0,24
Tolstoi Lev	1,00	0,92	0,96
Turgeniev Ivan	0,57	1,00	0,73

Результаты для классификатора, основанного на сжатии PPM, приведены в табл. 2.

Таблица 2

Точность, полнота и *F*-мера на текстах авторов XIX в. для PPM метода

Автор	Precision	Recall	F-measure
Chehov Anton	0,56	0,83	0,67
Dostoevskii Fedor	0,44	0,33	0,38
Gogol Nikolay	0,90	0,75	0,82
Goncharov Ivan	0,71	1,00	0,83
Kuprin Alexandr	0,40	0,33	0,36
Leskov Nikolai	0,71	0,83	0,77
Saltikov-Shedrin Mihail	0,67	0,17	0,27
Tolstoi Lev	1,00	0,92	0,96
Turgeniev Ivan	0,69	0,92	0,79

Для лучшего сравнения общую оценку для каждого классификатора приведем в табл. 3.

Таблица 3

Значения микро- и макропоказателей *F*-меры классификаторов на текстах авторов XIX в.

Метод	F-мера (micro)	F-мера(macro)
R-мера	0,75	0,71
PPM	0,68	0,65

В целом классификатор на основе R-меры показывает лучшие результаты для этой выборки, чем классификатор с использованием сжатия PPM. Лишь на некоторых классах PPM-классификатор показывает немного лучшие результаты, что можно связать с присутствием в алгоритме возможности учесть количественное вхождение одинаковых подстрок, за счет чего происходит более плотное сжатие текста.

Относительно второй выборки, основанной на текстах 21 класса, для усеченной *R*-меры были получены следующие характеристики, приведенные в табл. 4.

Таблица 4

Точность, полнота и *F*-мера на текстах авторов XX в. для *R*-меры

Фамилия автора	Precision	Recall	F-measure
Agafonov	0,88	1,00	0,93
Aristov	1,00	1,00	1,00
Azarov	0,96	1,00	0,98
Baganov	0,79	1,00	0,88
Belkin	1,00	1,00	1,00
BelobrovPopov	1,00	1,00	1,00
Belomlinskaya	1,00	1,00	1,00
Belov	1,00	0,78	0,88
Bonch	1,00	0,27	0,43
Bronin	1,00	1,00	1,00
Burmistrov	0,50	1,00	0,67
Galkin	1,00	0,29	0,45
Gergenreder	1,00	1,00	1,00
Glushkin	1,00	1,00	1,00
Svetlana	1,00	1,00	1,00
Velboi	0,10	0,20	0,13
Vershovskii	1,00	0,19	0,32
Veter	0,57	1,00	0,73
Vitkovskii	1,00	0,33	0,50
Voronov	1,00	1,00	1,00
Vulf	1,00	1,00	1,00

Для классификатора, основанного на сжатии PPM, результаты также приведены в табл. 5.

Таблица 5

Точность, полнота и *F*-мера на текстах авторов XX в. для PPM метода

Фамилия автора	Precision	Recall	F-measure
Agafonov	0,88	1,00	0,93
Aristov	0,96	1,00	0,98
Azarov	0,89	0,89	0,89
Baganov	0,96	0,88	0,92
Belkin	0,91	1,00	0,95
BelobrovPopov	1,00	1,00	1,00
Belomlinskaya	0,71	1,00	0,83
Belov	1,00	0,89	0,94
Bonch	0,91	0,91	0,91
Bronin	1,00	0,96	0,98
Burmistrov	0,48	0,94	0,64
Galkin	1,00	0,82	0,90
Gergenreder	0,95	1,00	0,98
Glushkin	1,00	1,00	1,00
Svetlana	1,00	1,00	1,00
Velboi	0,43	1,00	0,61
Vershovskii	1,00	0,19	0,32
Veter	0,95	1,00	0,98
Vitkovskii	1,00	0,33	0,50
Voronov	1,00	1,00	1,00
Vulf	1,00	1,00	1,00

Общая оценка для каждого классификатора на данной выборке представлена в табл. 6.

Таблица 6

Значения микро- и макропоказателей *F*-меры классификаторов на текстах авторов XIX в.

Метод	F-мера (micro)	F-мера(macro)
R-мера	0,85	0,80
PPM	0,90	0,87

На данной выборке классификация с помощью сжатия PPM показала себя несколько лучше – в выборке присутствует всего два класса, с которыми классификатор имеет проблемы, при этом только один класс вызывает у обоих классификаторов сбои. Успех классификации с помощью PPM метода на данной выборке достигнут во многом за счёт использования частотной характеристики символов в тексте для каждого конкретного класса.

Заключение

Оба классификатора на данных большинства рассмотренных примеров показали себя неплохо, но для каждого использованного метода стоит отметить факторы, которые снижают качество классификации. Метод на основе сжатия PPM крайне зависим от выбора модели и алгоритма PPM, использование новых его модификаций может увеличить процент успешного распознавания класса. Кроме того, данные тесты показали, что упор, сделанный на одну лишь частотную характеристику последовательностей символов в тексте в PPM методе, нередко увеличивает ошибку классификации при определении авторского стиля. В противоположность этому использование *R*-меры, как показало тестирование классификаторов на второй выборке, не столь эффективно, так как в ней используется информация лишь о присутствии подстроки в тексте и никак не учитывается частота данной подстроки в тексте. Использование частотной характеристики подстрок текста в развитии *R*-меры могло бы несколько улучшить качество классификации этим методом.

ЛИТЕРАТУРА

1. *Christopher M.B.* Pattern Recognition and Machine Learning // Springer Science. 2006.

2. *Khmelev D.V., Teahan W.J.* Verification of text collections for text categorization and natural language processing // Technical Report AIIA 03.1. School of Informatics, University of Wales. Bangor, 2003.
3. *Humnisett D., Teahan W.J.* Context-based methods for text categorization // Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR). The University of Sheffield, UK, 2004.
4. *Shevelyov O.G., Poddubny V.V.* Complex investigation of texts with the system "StyleAnalyzer" // Text and Lanquage / ed. by P. Grzyber, E. Kelih, J. Macutek. Wien : Praesens Verlag, 2010. P. 207–212.
5. *Шевелев О.Г.* Методы автоматической классификации текстов на естественном языке : учеб. пособие. Томск : ТМЛ-Пресс, 2007. 144 с.
6. *Хмелёв Д.В.* Классификация и разметка текстов с использованием методов сжатия данных. Краткое введение. 2003. URL: http://compression.graphicon.ru/download/articIes/classif/intro.html
7. *Поддубный В.В., Шевелев О.Г., Кравцова А.С., Фатыхов А.А.* Словарно-аналитический блок системы «Стилеанализатор» // Научное творчество молодежи : материалы XIV Всерос. науч.-практ. конф. (15–16 апреля 2010 г.). Томск : Изд-во Том. ун-та, 2010. Ч. 1. С. 138–140.

Ашуров Михаил Фаритович. E-mail: therevenge.amf@gmail.com
Томский государственный университет
Поступила в редакцию 4 сентября 2014 г.

Ashurov Mihail F. (Tomsk State University, Russian Federation).
Comparison of stream-based fiction text classification methods based on data compression and counting substrings.
Keywords: text classification, fiction classification, *R*-measure, PPM method, *F*-measure.

We consider the problem of comparing the quality of the natural language text classification based on the stream methods using *R*-measure and PPM compression. The task of text classification requires a certain amount of texts for each class (genre, author, etc.) in the classifier training. The process of classifier training is reduced to a concatenation of all texts in supertext by each class. In stream-based text classification X is a sequence (stream) of n elements (characters, words, phrases, etc.) [a]. A matching of the test text to any class, in cases of *R*-measure, is performed through maximum measure of closeness to supertext of a class. In the PPM method, the matching is determined by the smallest difference in lengths of a compressed concatenation of supertext and the test text by the length of the compressed supertext.

The truncated *R*-measure [a] counts all the test text (length n) substrings with the lengths from k_1 to k_2 in the supertext (unlike the basic method with $k_1 = 1$ and $k_2 = n$):

$$R(X \mid S) = r(X \mid S)/N\ ,\ r\left(X \mid S\right) = \sum_{k=k_1}^{k_2} c_k\left(X \mid S\right),\ N = \left(2(n+1) - (k_1 + k_2)\right)\left(k_2 - k_1 + 1\right)/2\ ,$$

$$c_k\left(X \mid S\right) = \sum_{i=k}^{n} c\left(x_{i-k+1} \ldots x_n \mid S\right),\ c\left(x_{i-k+1} \ldots x_n \mid S\right) = \begin{cases} 1,\ x_{i-k+1} \ldots x_n \subset S, \\ 0,\ x_{i-k+1} \ldots x_n \not\subset S. \end{cases}$$

where X is the test text, S is the supertext of the examined class, k is the length of the search substring, N is the number of substrings.

The PPM method is a context-dependent modeling with limited order, which allows to estimate the probability of a symbol appearance depending on the previous symbols [b]. The probability estimate using a context of length N is a context-limited model of order N (order-N, O-N). A good probability estimation of symbol appearance requires to consider contexts with different lengths. The calculated estimates are encoded by any entropy encoder, allowing to compress a text. Initially, a classification algorithm obtains the class supertext compressed by PPM method, then the concatenation of supertext, and finely the test text should be compressed in the same manner.

Both classifiers were tested by two text specific samples: the Russian prose texts of the 19th century and the 90s of the 20th century. The first sample contains only 9 authors representing the classes, the second sample - 21 authors. Each sample contains about 100 texts that make a training sample. The supertexts formed by each class texts were normalized by volume. Test samples were obtained from texts that were out of training sample, except two texts for each class, and they were normalized by the number and volume (about 100 thousand characters for a test text).

To test the stream classification quality software modules that allow classify texts by R-measure and PPM method have been designed and implemented in C#. The problem of the k_1 and k_2 parameter selection in R-measure has been solved by natural language features. The minimum length of substrings (k_1) is equal to 10 characters because a shorter substrings can match with a lot of words in Russian language, which are common for all authors, and that can reduce stylistic features detection of different authors. The maximum length of the substrings (k_2) is 45 characters that allows to process even a phrase with 3 or 4 words. Using a greater length of substring seems not so useful because such long phrase cannot be repeated many times by various authors. Note that the problem of plagiarism is completely excluded from our consideration. Classifier Module based on the PPM method uses the PPMD algorithm of the order 5.

Classification quality for each class is estimated by F-measure, namely, the harmonic mean of precision and recall [a]. The mean of F-measure values by all classes is taken as the estimate of the stream-based classification quality in general.

The classifier quality characteristics for each class are obtained for the test samples. The means of F-measure for each classifier are calculated and the conclusions about the classification quality are made. The factors that reduce the classification quality were described for both methods.

---

[a] Shevelyov O.G. Automatic natural language text classification methods: Tutorial. Tomsk: TML-Press, 2007. 144 p.

[b] Khmelev D. V. Text classification and markup using compression methods. Introduction, 2003. URL: http://compression.graphicon.ru/download/articIes/classif/intro.html

REFERENCES

1. Bishop C.B. *Pattern Recognition and Machine Learning*. Springer Science, 2006.
2. Khmelev D.V., Teahan W.J. Verification of text collections for text categorization and natural language processing. *Technical Report AIIA 03.1.* School of Informatics, University of Wales. Bangor, 2003.
3. Humnisett D., Teahan W.J. Context-based methods for text categorization. *Proceedings of the 27th Annual International ACM SIGIR Conference (SIGIR),* The University of Sheffield, UK, 2004. DOI: 10.1145/1008992.1009129
4. Shevelyov O.G., Poddubny V.V. *Complex investigation of texts with the system "StyleAnalyzer"*. In: Grzyber P., Kelih E., Macutek J. (eds.) *Text and Lanquage*. Wien: Praesens Verlag, 2010, pp. 207-212.
5. Shevelev O.G. *Metody avtomaticheskoy klassifikatsii tekstov na estestvennom yazyke* [Automatic natural language text classification methods]. Tomsk: TML-Press Publ., 2007. 144 p. (in Russian).
6. Khmelev D.V. *Klassifikatsiya i razmetka tekstov s ispol'zovaniem metodov szhatiya dannykh* [Text classification and text markup using compression methods]. Available at: http://compression.graphicon.ru/download/articIes/classif/intro.html.
7. Poddubny V.V., Shevelyov O.G., Kravtsova A.S., Fatihov A.A. [Dictionary-analysis unit of the system “StyleAnalyzer”. *Nauchnoe tvorchestvo molodezhi : materialy XIV Vseros. nauch.-prakt. konf* [Scientific creativity of the youth. Proc. of the 14th All-Russian Scientific. Pract. conf.]. Tomsk: Tomsk State University Publ., 2010, pt. 1, pp. 138-140. (In Russian).