Классификация научных текстов по отраслям знаний *

Сухарева А. В., Царьков С. В.
suhareva_anzhela@mail.ru
Московский физико-технический институт

В этой статье рассматривается задача классификации публикаций, относящихся к различным предметным областям. Классификация выполняется на основе предметного словаря. Разработанный алгоритм применяется для оценивания и сравнения различных методов автоматического выделения терминов. Особенность исследования в том, что была сравнена униграммная классификация на всех словах с n-граммной классификацией. Проведенные нами эксперименты показали, что использование униграмм позволяет повысить качество классификации научных текстов. В тоже время, униграммная модель имеет ряд недостатков, которые устраняются в модели n-грамм.
Задача была решена с помощью линейного многоклассового классификатора (на основе наивного байесовского классификатора) с отбором признаков, имеющим линейное по числу объектов и числу признаков время обучения. В экспериментах на задаче классификации научных текстов его качество сравнивается с SVM и Logistic regression по униграммным и n-граммным признакам.

Ключевые слова: *наивный байесовский классификатор, AUC, MAUC, униграммная модель, n-граммная модель, отбор признаков.*

Classification of scientific texts by industry knowledge*

Sukhareva A. V.,[1]***, Tsarkov S. V.***[2]
Moscow Institute of Physics and Technology

In this article we consider the problem of classification of publications that relate to different subject areas. The classification is performed using the dictionary. The algorithm will be used for evaluation and comparison of different methods for automatic selection of terms. The study compared unigram classification on all words with n-gram classification. It is shown that the use of n-grams can improve the quality of classification of scientific texts.
The problem was solved with a linear multi-class classifier (based on Naive Bayes classifier) with feature selection, having linearly on the number of objects and the number of signs of the training. In experiments on the problem of classifying the quality of scientific texts compared with SVM and Logistic regression on unigram and n-gram features.

Keywords: *Naive Bayes classifier, AUC, MAUC, unigram model, n-gram model, feature selection.*

Введение

В настоящее время наблюдается всплеск научных работ, посвященных рубрикации текстов на основе методов машинного обучения [1, 2, 3]. Классификация документов используется, например, в электронных библиотеках научных публикаций для автоматического заполнения метаописаний при поступлении новых документов в библиотеку. Одна из таких коллекций публикаций рассматривается в качестве выборки. Применение методов

Работа выполнена при финансовой поддержке РФФИ, проект РФФИ: 14-07-31176. Научный руководитель: Воронцов К. В. Консультант: Царьков С. В.

машинного обучения для классификации текстов очень эффективно при наличии качественно размеченной обучающей коллекции. В докладах конференции [4] было отмечено, что для больших рубрикаторов (более 500 рубрик) из-за трудности формирования непротиворечивой обучающей выборки единственный работающий подход — трудоемкое ручное описание смысла каждой рубрики. Таким образом, задача разработки эффективных алгоритмов классификации является актуальной.

Рассмотрим один из самых популярных и старейших подхов к классификации — байесовский подход. Байесовский классификатор [5] позволяет определить вероятность принадлежности объекта к одному из классов. При этом выдвигается предположение о независимости влияния на эту вероятность различных атрибутов объектов — так называемое предположение об условной независимости классов, которое существенно упрощает сопутствующие вычисления. Байесовский классификатор относит объект к определенному классу тогда и только тогда, когда апостериорная вероятность принадлежности объекта к этому классу больше апостериорной вероятности принадлежности объекта к любому другому классу.

Байесовский подход к классификации основан на теореме [6, 7], утверждающей, что если плотности распределения каждого из классов известны, то искомый алгоритм можно выписать в явном аналитическом виде. Более того, этот алгоритм оптимален, то есть обладает минимальной вероятностью ошибок.

На практике плотности распределения классов, как правило, не известны. Их приходится оценивать (восстанавливать) по обучающей выборке. В результате байесовский алгоритм перестаёт быть оптимальным, так как восстановить плотность по выборке можно только с некоторой погрешностью. Чем короче выборка, тем выше шансы подогнать распределение под конкретные данные и столкнуться с эффектом переобучения.

Байесовский подход к классификации лежит в основе достаточно удачных алгоритмов классификации. Одним из них является наивный байесовский классификатор.

Наивный байесовский классификатор [6, 7, 8] — простой вероятностный классификатор, основанный на применении Теоремы Байеса со строгими (наивными) предположениями о независимости. В зависимости от точной природы вероятностной модели, наивные байесовские классификаторы могут обучаться очень эффективно. Во многих практических приложениях, для оценки параметров для наивных байесовых моделей используют метод максимального правдоподобия; другими словами, можно работать с наивной байесовской моделью, не веря в байесовскую вероятность и не используя байесовские методы.

Преимуществами наивного байесовского классификатора являются: малое количество данных для обучения, необходимых для оценки параметров, простота реализации и низкие вычислительные затраты при обучении и классификации. В тех редких случаях, когда признаки действительно независимы (или почти независимы), наивный байесовский классификатор (почти) оптимален. Основной его недостаток — относительно низкое качество классификации в большинстве реальных задач.

Результат классификации зависит не только от выбора алгоритма, но и от того, какой набор характеристик используется для составления вектора признаков. Наиболее распространенный способ представления документа в задачах компьютерной лингвистики и поиска — это униграммы и n-граммы. В данной работе исследуется качество классификации в зависимости от применения униграмм и n-грамм в качестве признакового описания документов.

Описание данных

Выборка — коллекция из 7852 авторефератов диссертаций, которые представляют собой тексты на русском языке. Половина публикаций была отнесена к обучающей выборке (4233 автореферата), другая половина — к тестовой (3619 автореферата). В задаче рассматривается 20 классов документов. Классами являются области науки. Распределение документов по классам представлено в табл.1, из которой видно, что данная задача является задачей с несбалансированными классами (unbalanced classes), поскольку некоторые классы встречаются гораздо чаще, чем остальные.

Таблица 1. Статистика по данным.

Отрасль наук	**Обучающая выборка**		**Тестовая выборка**	
	кол-во авторефератов	доля в выборке, %	кол-во авторефератов	доля в выборке, %
медицинские науки	887	20,954	885	24,454
технические науки	614	14,505	577	15,944
экономические науки	578	13,655	319	8,815
физико-математические науки	321	7,583	305	8,428
биологические науки	305	7,205	242	6,687
педагогические науки	243	5,741	226	6,245
филологические науки	227	5,363	195	5,388
исторические науки	197	4,654	148	4,090
философские науки	148	3,496	136	3,758
юридические науки	137	3,236	124	3,426
сельскохозяйственные науки	124	2,929	111	3,067
химические науки	113	2,670	61	1,686
социологические науки	61	1,441	59	1,630
политические науки	60	1,417	54	1,492
психологические науки	54	1,276	48	1,326
геолого-минералогические науки	48	1,134	33	0,912
искусствоведение	33	0,780	31	0,857
географические науки	32	0,756	30	0,829
ветеринарные науки	31	0,732	19	0,525
культурология	20	0,472	16	0,442

В данной работе сравниваются два вида признаков: униграммы и n-граммы. На вход алгоритма подаются словари коллекции текстов, один из них состоит из 139229 униграмм, другой содержит 197191 n-грамм. Униграммы были полученны после следующей обработки:

— слова текстов лемматизированы;
— удалены стоп-слова и служебные части речи;
— удалены слова, встречающиеся менее двух раз в документе.

Для автоматического выделения n-грамм в тематических моделях коллекций текстовых документов использовался двухэтапный алгоритм [9]. На первом этапе формируется из-

быточно большой лексикон из n-грамм, отобранных по морфологическим признакам и статистическим критериям релевантности и устойчивости n-грамм. На втором этапе отбираются n-граммы, частота встречаемости которых не менее 6.

Постановка задачи

Разработать линейный многоклассовый классификатор с отбором признаков, имеющий линейное по числу объектов и числу признаков время обучения. В качестве признаков используются частоты слов и словосочетаний. В экспериментах на задаче классификации научных текстов его качество сравнивается с SVM и Logistic regression по униграммным и n-граммным признакам.

Введем следующие обозначения:

X – коллекция текстовых документов, состоящая из документов x;

W – словарь коллекции текстов, состоящий из n-грамм w;

Y – конечное множество классов, состоящее из классов y.

Предполагается, что документы $x \in X$ описываются бинарными признаками $(x^1, \ldots, x^N)$:

$$b_w(x) = [f_w(x) \geqslant 0], \tag{1}$$

где $b_w(x) \in \{0, 1\}$ — бинарный признак, $f_w(x)$ — частота встречаемости n-граммы w в документе x.

Задача многоклассовая. Для выбора других классов будем использовать стратегии каждый-против-каждого. Заметим, что для каждого класса важна только небольшая часть признаков, которые позволяют отнести данный документ x к некоторому классу y. Отбор информативных терминов-признаков для каждого класса из множества Y заключается в том, чтобы отсортировать n-граммы по убыванию абсолютного значения весов n-грамм и взять первые K. K — настраиваемый параметр модели, причем свой для каждого класса. Для определения веса термина зафиксируем один класс y. Вычисляем для каждой n-граммы w $tf(w, y)$ и $tf(w, \overline{y})$ в соответствии со стратегией каждый-против-каждого:

$$tf(w, y) = \frac{\sum\limits_{x \in Y \bigcap w \in x} x}{\sum\limits_{x \in Y} x}, \tag{2}$$

где $tf(w, y)$ — доля документов класса y с признаком w.

$$tf(w, \overline{y}) = \frac{\sum\limits_{x \in \overline{Y} \bigcap w \in x} x}{\sum\limits_{x \in \overline{Y}} x}, \tag{3}$$

где $tf(w, \overline{y})$ — доля документов фиксированного класса $\overline{y}$ с признаком w.

Определяем вес термина w в классе y $wt(w, y)$:

$$wt(w, y) = \sqrt{tf(w, y)} - \sqrt{tf(w, \overline{y})} \tag{4}$$

Требуется построить эмпирическую оценку плотности распределения, приближающую неизвестную плотность вероятностного распределения, сгенерировавшего обучающую выборку X^l.

Гипотеза 1. Признаки $x^1, \ldots, x^n$ являются независимыми случайными величинами. Следовательно, функции правдоподобия классов представимы в ввиде

$$p_y(x) = p_{y^1}(x^1) \cdots p_{y^n}(x^n), \tag{5}$$

где $p_{y^j}(x^j)$ — плотность распределения значений j-го признака для класса y.

Алгоритм восстанвления плотности распределения на основе гипотезы 1. называется наивным байесовским классификатором [6, 7]. В работе наивный байесовский классификатор используется для построения линейного многоклассового классификатора:

$$a(x) = \arg\max_y \sum_{w=1}^{K} wt(w, y) x^w, x \in X, y \in Y, \tag{6}$$

где $K = \{k_1, \ldots, k_m\}$ — число информативных признаков, $m = \frac{c(c+1)}{2}$ — число бинарных классификаторов, c — количество классов, $wt(w, y)$ — вес признака w в классе y, x^w — признак объекта x с истинным номером w, который существовал у признака w до сортировки признаков.

В качестве функционала качества, по которому будет вестись сравнение моделей, используется $MAUC$ [10] (мультиклассовый AUC):

$$MAUC = \frac{1}{c} \sum_{i=1}^{c} \frac{1}{c-1} \sum_{j=1, j \neq i}^{c} A_{ij}, \tag{7}$$

где c — количество классов, $y_i \in \{1, \ldots, c\}$, A_{ij} — это AUC [6, 7] между классами i и j .

Описание алгоритма

Ставится задача восстановления зависимости $y = f(x)$ по точкам обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$.

Дано пространство объектов $X = \mathbb{R}^n$, множество классов $Y = \{1, \ldots, c\}$.

Необходимо найти по обучающей выборке X^l параметр $K = \{k_1, \ldots, k_m\}$ — число информтивных признаков алгоритма классификации (6).

В качестве критериев качества используются AUC и $MAUC$. На выходе алгоритма получаем: AUC, $MAUC$ на обучении и контоле в зависимости от параметра K. В модели предполагается, что n-грамма есть, если она встречается в тексте хотя бы 6 раз (для униграммной модели порог 2). Такой порог позволяет вычислить матрицу объекты-признаки за разумное время.

Исходные документы хранятся в виде «мешка слов». Операцией объединения слов публикаций получаем словарь классиификации. С помощью словаря по формуле (1) документы преобразуются в бинарные признаки. В результате векторизации документов получим матрицу «объекты-признаки»: столбцы — признаки (n-граммы), строки — объекты (документы публикаций).

По обучающей выборке X^l вычисляем матрицу весов признаков: вычислить веса $wt(x^j, y)$ по формуле (4) для всех признаков $j = 1 \ldots n$ для каждого класса $y \in Y$. Сортируем признаки по убыванию $|wt(x^j, y)|$. Следующие шаги выполняем для обучающей выборки.

Решение многоклассовой задачи сводится к решению задачи бинарной классификации для каждой пары классов, причем порядок классов в паре не важен в силу симметрии задачи относительно критерия качества AUC. Всего получается $m = \frac{c(c+1)}{2}$, где c — число

Алгоритм 1 Построение ROC-кривой и вычисление AUC за $O(l)$.

Вход: выборка X^l, функция $SCORE(x, wt(w, y)) = \sum_{w=1}^{K} wt(w, y)x^w, x \in X, y \in Y$;

Выход: $\{(FPR_i, TPR_i)\}_{i=0}^{l}$, AUC — площадь под ROC-кривой;

```
l_0 := Σ_{i=1}^{l} [y_i = 0]; l_1 := Σ_{i=1}^{l} [y_i = 1];
упорядочить выборку по убыванию SCORE(x, wt(w, y));
(FPR_i, TPR_i) := (0, 0); AUC := 0;
для i := 1, ..., l
    если y_i = 0 то
        сместиться на один шаг вправо:
        FTR_i := FPR_{i-1} + 1/l_0; TPR_i := TPR_{i-1};
        AUC := AUC + 1/l_0 TPR_i;
    иначе
        сместиться на один шаг вверх:
        TPR_i := TPR_{i-1} + 1/l_1; FPR_i := FPR_{i-1};
```

классов, моделей бинарной классификации. Подбираем параметр k_i для каждой модели $i \in \{1, \ldots, m\}$. Вначале положим $k_i = 1$. Вычисляем для каждой модели критерий качества классификации AUC [11] по алгоритму 1.

Вычисляем критерий $MAUC$ по формуле (5) при каждом зачении k_i. Запоминаем оптимальный k_i. Из моделей получаем итоговую модель. По итоговой модели на тестовой выборке рассчитываем $MAUC$. Параметр $K = \{k_1, ..., k_m\}$ итоговой модели — число уникальных признаков среди всех отобранных информативных признаков бинарных моделей. Аналогично вычисляем для контрольной выбороки критерий качества AUC и $MAUC$. На контольной выборке, также как и на обучающей, используем обученные веса и ответы, чтобы проверить критерий качества классификации AUC.

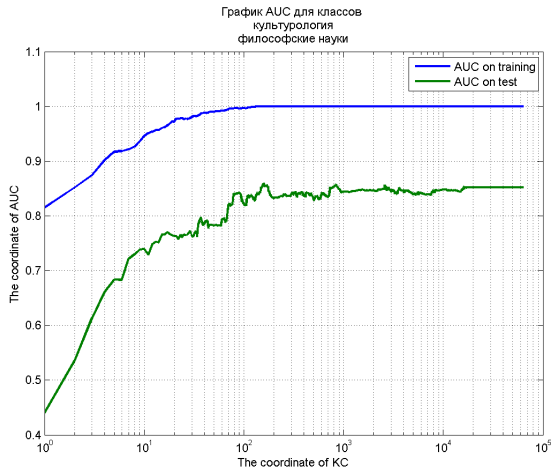
Вычислительный эксперимент

В ходе экспериментов были рассмотрены униграммная и n-граммная модели признакового описания объектов. В качестве стратегии разбиения на классы использовалась стратегия каждый-против-каждого. Вычислялся AUC и $MAUC$ на обучении и на контроле в зависимлсти от числа признаков K.

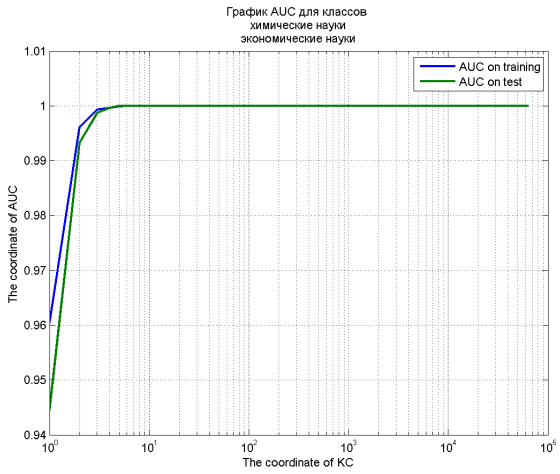
На многих графиках AUC униграммой модели видно, что если брать первые 10 тысяч признаков и более, качество ухудшается (см. Рис.3). Это связано с тем, что слишком много признаков на обучении являются информативными, и в имеющемся объёме данных просто не хватает надёжной статистики, чтобы совсем хорошо отсортировать признаки. В результате мы наблюдаем интересный эффект — как переобучается отбор признаков. Оказывается, что использовать первые 500 признаков — гораздо надёжнее, чем первые 10 тысяч. То есть при таком соотношении числа объектов и числа признаков наивный байесовский классификатор переобучается. При числе признаков меньшем 10 может наблюдаться сильное влияние шума (см. Рис.3(а)), из-за которого критерий качества на конторле лучше, чем на обучении.

В результате эксперимента было выявлено, что хуже всех различаются классы культурология и философские науки (см. Рис.1(а)), но применение n-грамм не позволило улучшить этот результат (см. Рис.2).

Эксперименты показали, что использование n-грамм в качестве признаков наивного байесовского классификатора приводит к ухудшению качества классификации (см. Рис.5), однако, отбор признаков не переобучается (см. Рис.3) и гораздо меньше влияние шума.

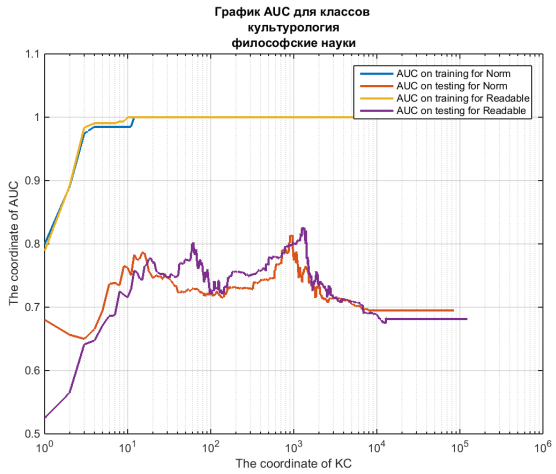


(a) Плохо различимые классы.

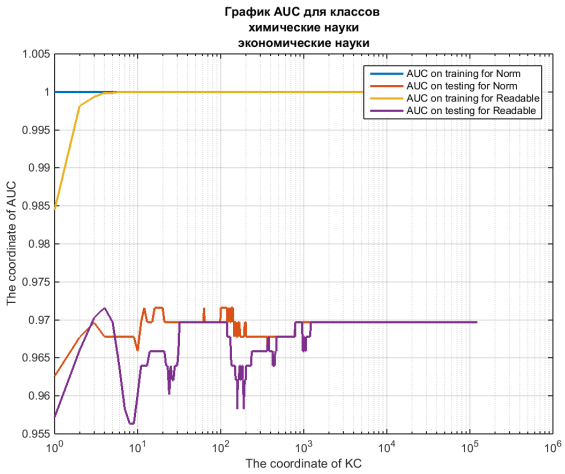


(b) Хорошо различимые классы.

Рис. 1. Степень различимости классов в униграммной модели.

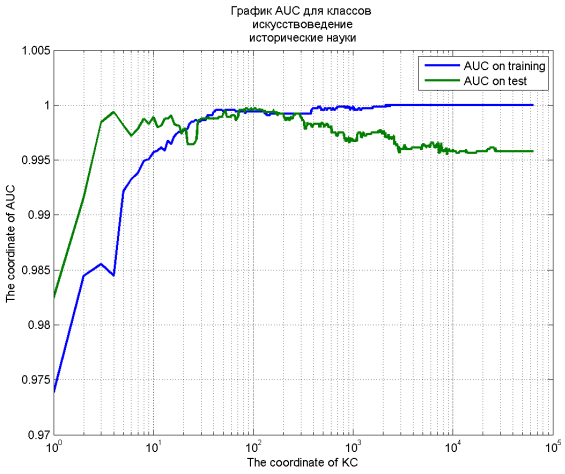


(a) Плохо различимые классы.

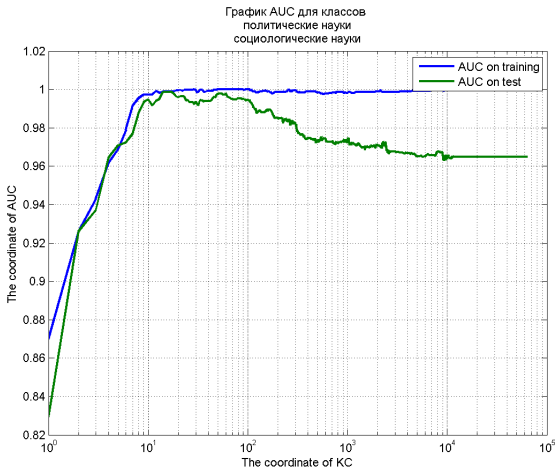


(b) Хорошо различимые классы.

Рис. 2. Степень различимости классов в n-граммной модели.

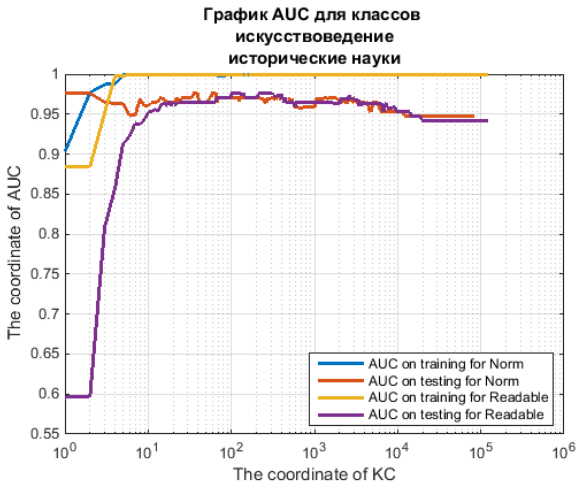


(a) Сильное влияние шума при небольшом числе признаков.

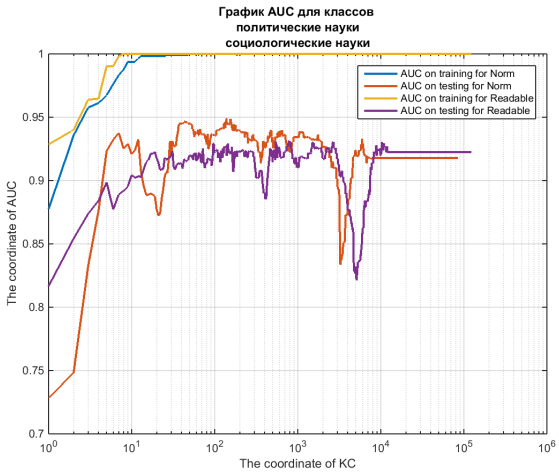


(b) Шум не влияет на критерий качества.

Рис. 3. Переобучение отбора признаков в униграммной модели.



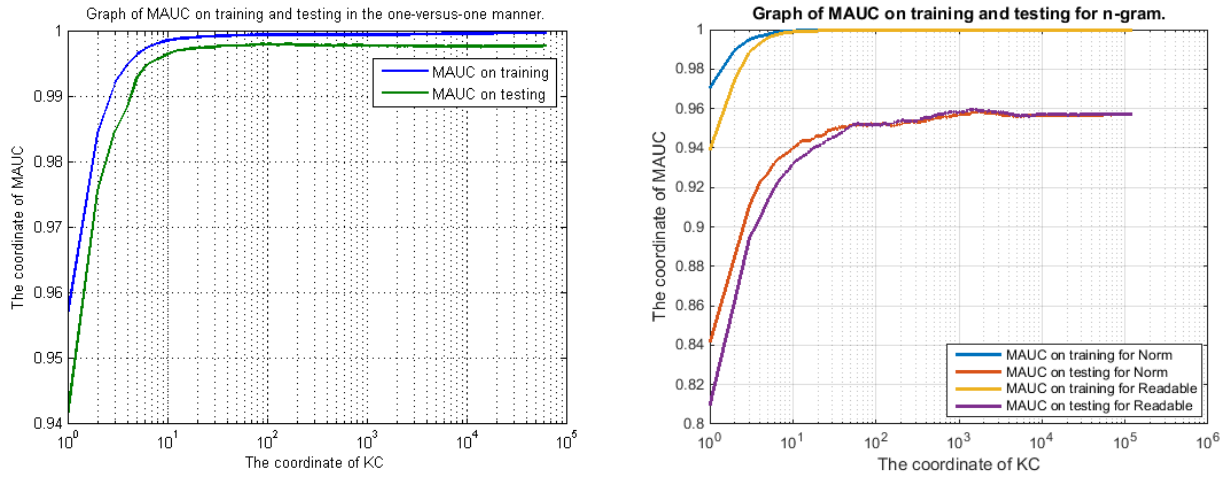
(a) Сильное влияние шума при небольшом числе признаков.



(b) Шум не влияет на критерий качества.

Рис. 4. Переобучение в n-граммной модели.

Качество классификации в среднем по всем классам показывает график $MAUC$ (см. Рис.5).



(a) Униграммная модель. (b) n-граммная модель.

Рис. 5. Мультиклассовый AUC.

Заключение

В работе представлены эксперименты по сравнению униграммной и n-граммной моделей классификации. Эксперименты, проведённые на русскоязычных авторефератах диссертаций, показывают, что униграммы дают лучшие результаты, чем n-граммы, по критерию качества $MAUC$. Модель «мешка слов», в которой каждый документ рассматривается как набор встречающихся в нем слов, имеет ряд недостатков. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов в документах друг от друга. На данный момент проведено множество исследований, посвященных изучению вопроса добавления словосочетаний, n-грамм и многословных терминов в тематические модели. Однако часто это приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели [12, 13, 14]. Комбинация униграмм и биграмм позволяет улучшить результаты.

В статье рассматривается линейный многоклассовый классификатор (на основе наивного байесовского классификатора) с отбором признаков, имеющим линейное по числу объектов и числу признаков время обучения. В дальнейшем планируется провести эксперименты по сравнению построенного классификатора с SVM и Logistic regression по униграммным и n-граммным признакам на задаче классификации научных текстов.

Литература

[1] Гринева М., Гринев М., and Лизоркин Д. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов. In *Труды Института системного программирования РАН*, 2009.

[2] Aizawa A. An information-theoretic perspective of tf–idf measures. *Information Processing and Management: an International Journal*, page 45 – 65, 2003.

[3] Браславский П. and Соколов Е. Сравнение пяти методов извлечения терминов произвольной длины. In *Компьютерная лингвистика и интеллектуальные технологии: ежегодная Международная конференция «Диалог» (Бекасово, 4–8 июня 2008 г.)*, Вып. 7 (14), pages 67–74, М.: РГГУ, 2008.

[4] Harding S.M. The INQUERY Retrieval System Callan J.P., Croft W.B. Proceedings of dexa-92, 3rd international conference on database and expert systems applications. pages 78–83, 1992.

[5] Pat Langley George H. John. Estimating continuous distributions in bayesian classifiers. *Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[6] К.В. Воронцов. Курс лекций К.В. Воронцова. Машинное обучение., 2011.

[7] Hastie T., Tibshirani R., and Friedman J. *The Elements of Statistical Learning*. Springer, 2001.

[8] Енюков И. С. Мешалкин Л. Д. Айвазян С. А., Бухштабер В. М. *Прикладная статистика: классификация и снижение размерности*. Финансы и статистика, М., 1989.

[9] Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов. *Естественные и технические науки №6(62).*, pages С. 456–464., 2012.

[10] D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Mach. Learning 45*, page 171–186, 2001.

[11] К.В. Воронцов. Задача диагностики заболеваний по электрокардиограмме, 2014.

[12] H. Wallach. Topic modeling: beyond bagofwords. In *In the Proceedings of the 23rd International Conference on Machine Learning*, page pp. 977–984, 2006.

[13] A. McCallum X. Wang and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *In the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, page pp. 697–702, 2007.

[14] . H. Wallach, pp. 977–984. Topic modeling: beyond bagofwords. *In the Proceedings of the 23rd International Conference on Machine Learning*, 2006.