Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Сухарева Анжелика Вячеславовна

Оценивание качества выделения терминов в задаче классификации текстовых документов

010990 Интеллектуальный анализ данных

магистерская диссертация

Научный руководитель:
д.ф.-м.н., проф. МФТИ
Воронцов Константин Вячеславович

Москва
2016 г.

Содержание

Аннотация

В работе рассматривается критерий качества, применяемый для оценивания и сравнения различных алгоритмов выделения терминов (Term Extraction). Критерий построен на основе качества классификации, которая показывает, насколько хорошо были отобраны термины предметных областей. В результате экспериментов было выявлено, что использование мультиграмм позволяет повысить качество классификации научных текстов.

С целью повышения чувствительности критерия к качеству выделения терминов была рассмотрена классификация фрагментов авторефератов. Задача была решена с помощью композиции линейных многоклассовых классификаторов (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 методом опорных векторов и тематической моделью классификации по униграммным и мультиграммным признакам.

1 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 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.

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. Документы хранятся в виде так называемого «мешка слов» («bag of words»). Униграммная модель — наиболее популярная модель представления текстовых документов, которая рассматривает каждый термин в качестве независимой случайной величины вне контекста и связи с другими словами текста. n-граммы получены с помощью алгоритма автоматического выделения ключевых фраз [5] по коллекции текстовых документов. Они образуют лексикон вероятностной тематической модели. В данной работе исследуется качество классификации в зависимости от применения униграмм и n-грамм как признакового описания документов.

Цель данного исследования: разработать способы измерения качества выделения терминов в задачах классификации текстов.

Проблемы исследования:

- Как качество выделения терминов влияет на качество классификации?
- Как построить чувствительный критерий качества выделения терминов?

Решение: строить как можно более точные модели классификации и, измеряя их качество, тем самым измерять качество мультиграммных словарей терминов.

Рассмотрим один из самых популярных и старейших подходов к классификации — байесовский подход. Байесовский классификатор [6] 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 [7, 8], 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 (Naïve Bayes, NB) [7, 8, 9] — 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.

Также NB применяется в качестве базового классификатора при построении композиции классификаторов. Ансамбли классификаторов являются эффективным методом повышения точности классификации, которое происходит за счет компенсации ошибок слабых классификаторов. Дальнейшее повышение качества классификации можно достичь с помощью тематической модели классификации, которая показывает хорошие результаты на больших текстовых коллекциях с большим числом несбалансированных, взаимозависимых классов [10].

2 Постановка задачи

2.1 Задача выделения терминов

Языковые статистические модели (Statistical Language Modelling) [11] являются моделями информационного поиска. Они применяются во многих областях автоматической обработки текста и речи, в частности, в распознавании речи, машинном переводе, морфологическом и синтаксическом анализе текста.

В большинстве случаев используется униграммная языковая модель, в которой терминами w считаются отдельные слова. Модель униграмм не учитывает связи между словами, так как считает все слова независимыми. Кроме того известно, что униграммная модель способна порождать ложные связи между документами [12]. Более эффективным способом представления данных является n-граммная модель, которая выделяет словосочетания, содержащие гораздо больше информации о тематике текста. В работе показано, что использование n-грамм (фраз, состоящих из n слов, в том числе и одного слова) позволяет повысить качество классификации при допустимом увеличении признакового пространства. Однако большие n приводят к большей разреженности обучающего корпуса, что является некоторой проблемой.

Задача выделения терминов (Term Extraction) заключается в том, чтобы по коллекции текстовых документов сформировать лексикон (словарь терминов) коллекции. В работе исследуются ключевые фразы, автоматически построенные одним из алгоритмов Term Extraction [5]. Алгоритм состоит из двух этапов.

На первом этапе формируется избыточно большой лексикон из n-грамм, отобранных по морфологическим признакам и статистическим критериям релевантно-

сти и устойчивости n-грамм. Число n-грамм, содержащихся в коллекции, может быть огромным, но не все они являются ключевыми.

На втором (статистическом) этапе отбираются n-граммы, наиболее полезные для тематической модели, что позволяет существенно сократить лексикон без ухудшения качества тематической модели. Для отбора ключевых фраз использовалась технология автоматического выделения ключевых фраз для тематического моделирования без привлечения внешней информации. Здесь может быть один из методов: *TF-IDF*, *Termhood*. Нужно определить, какой метод лучше выделяет термины.

2.2 Формальная постановка задачи

Разработать линейный многоклассовый классификатор с отбором признаков (композицию линейных многоклассовых классификаторов), имеющий линейное по числу объектов и числу признаков время обучения. В качестве признаков используются частоты униграмм и n-грамм, отфильтрованные по порогу. Исследование проводится на двух типах коллекций: коллекция, полученная по целым текстам авторефератов, и коллекции, в которых каждый автореферат разрезался по строкам на 25 одинаковых частей.

Обозначим $X \in \mathbb{R}^n$ — коллекцию текстовых документов, состоящую из документов x, а $Y = \{1, \ldots, C\}$ — конечное множество классов. Данные представлены в виде «мешка слов». Предполагается, что документы $x \in X$ описываются бинарными признаками $(x^1, \ldots, x^n)$:

$$b_w(x) = [f_w(x) \geqslant th], \tag{2.1}$$

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

Ставится задача восстановления зависимости $y = f(x)$ по точкам обучающей выборки $X^l = (x_i, y_i)_{i=1}^l$. Отметим, что объекты описываются зависимыми признаками и число признаков много больше числа объектов $l \ll n$, также задача является задачей с несбалансированными классами.

Замечание 2.1. Для каждого класса важна только небольшая часть признаков, которые позволяют отнести данный документ x к некоторому классу y. Такие признаки называются информативными.

Итак, необходимо найти по обучающей выборке X^l:

Алгоритм 2.1. Построение ROC-кривой и вычисление AUC за $O(l)$.

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

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

$l_0 := \sum_{i=1}^{l} [y_i = 0]$; $l_1 := \sum_{i=1}^{l} [y_i = 1]$;
упорядочить выборку по убыванию $SCORE(x, w_y)$;
$(FPR_i, TPR_i) := (0, 0)$; $AUC := 0$;
для $i := 1, \ldots, l$
  если $y_i = 0$ **то**
    сместиться на один шаг вправо:
    $FTR_i := FPR_{i-1} + \frac{1}{l_0}$; $TPR_i := TPR_{i-1}$;
    $AUC := AUC + \frac{1}{l_0} TPR_i$;
  иначе
    сместиться на один шаг вверх:
    $TPR_i := TPR_{i-1} + \frac{1}{l_1}$; $FPR_i := FPR_{i-1}$;

1. матрицу весов признаков $W = w_y^j$, w_y^j — вес j признака в классе y;

2. параметр $K = \{k_1, \ldots, k_m\}$ — число информативных признаков алгоритма классификации (2.3);

3. количество выделенных терминов l^*, доставляющее максимум критерия качества классификации AUC.

На выходе алгоритма получаем: AUC [7, 8] на контроле в зависимости от параметра K. В качестве функционала качества, по которому будет вестись сравнение моделей, используется AUC по алгоритму 2.1, $MAUC$ [13]:

$$MAUC = \frac{2}{C(C-1)} \sum_{i<j} \frac{AUC_{ij} + AUC_{ji}}{2}, \tag{2.2}$$

где C — число классов. Так как признаки бинарные, то $AUC_{ij} = AUC_{ji}$.

Задача многоклассовая. Для выбора других классов будем использовать различные многоклассовые стратегии. Требуется построить эмпирическую оценку плотности распределения, приближающую неизвестную плотность вероятностного распределения, сгенерировавшего обучающую выборку X^l. Простой алгоритм восстановления плотности распределения на основе гипотезы независимости признаков называется наивным байесовским классификатором [7, 8]. В работе наивный байесовский

классификатор используется для построения линейного многоклассового классификатора:

$$a(x, w) = \arg\max_y \sum_{j=1}^{K} w_y^j x^j,\ x \in X,\ y \in Y, \tag{2.3}$$

где $K = \{k_1, \ldots, k_m\}$ — число информативных признаков, m — число бинарных классификаторов, w_y^j — вес признака j в классе y, x^j — признак объекта x с номером j.

3 Наивный байесовский классификатор

3.1 Байесовские методы классификации

Рассмотрим вероятностную постановку [7] задачи классификации. Пусть имеется простая выборка $X^l = (x_i, y_i)_{i=1}^l$ из неизвестного распределения

$$p(x, y) = P(y)p(x|y),$$

где $P(y)$ — вероятности появления объектов каждого из классов, $p(x|y)$ — функции правдоподобия классов. Требуется построить эмпирические оценки априорных вероятностей $\hat{P}(y)$ и функций правдоподобия $\hat{p}(x|y)$ для каждого из классов $y \in Y$.

Теорема 3.1. *Если $P(y)$ и $p(x|y)$ известны, а классы равнозначны, то минимум среднего риска достигается алгоритмом [7, 14]*

$$a(x) = \arg\max_{y \in Y} P(y)p(x|y). \tag{3.1}$$

Предполагается, что обучающие объекты выбираются случайно и независимо друг от друга. В этом случае эмпирические оценки априорных вероятностей

$$\hat{P}(y) = \frac{1}{l} \sum_{i=1}^{l} [y_i = y]$$

являются несмещенными.

Согласно параметрическому подходу к классификации функции правдоподобия считаются известными с точностью до параметров θ_y

$$p(x|y) = p(x; \theta_y).$$

Пусть $X_y = \{(x_i, y_i)_{=1}^l |\, y_i = y\}$ подвыборка выборки X^l объектов класса y. Поскольку выборка X_y простая, несмещенные оценки параметров θ_y строятся по выборке, применяя принцип максимума правдоподобия:

$$\ln \prod_{i=1}^{l} p(x_i, y_i) = \sum_{y \in Y} \sum_{x_i \in X_y} \ln p(x; \theta_y) \to \max_{\theta_y}. \tag{3.2}$$

3.2 Наивный линейный байесовский классификатор

Пусть объекты $x \in X$ описываются n бинарными признаками $(x^1, \ldots, x^n)$. Наивный байесовский классификатор основан на сильном вероятностном предположении о том, что признаки являются независимыми случайными величинами.

Гипотеза 3.1. *Признаки $x^1, \ldots, x^n$ являются независимыми случайными величинами. Следовательно, функции правдоподобия классов представимы в виде*

$$p(x|y) = p(x^1, x^2, \ldots, x^n|y) = p(x^1|y) \cdots p(x^n|y),$$

где $p(x^j|y) = p(x^j; \theta_y^j)$ — плотность распределения значений j-го признака для класса y.

Согласно гипотезе 3.1 принцип максимума логарифма правдоподобия (3.2) принимает вид:

$$\sum_{j=1}^{n} \sum_{y \in Y} \sum_{x_i \in X_y} \ln p(x^j; \theta_y^j) \to \max_{\theta_y^j}. \tag{3.3}$$

Сильное предположение независимости признаков редко выполняется на практике, тем не менее, NB очень популярен из-за простоты реализации и достаточно высокого качества классификации во многих прикладных задачах. Кроме того, в работе NB использовался в качестве элементарного «строительного блока» при построении композиции классификаторов.

Без ограничения общности, пусть имеется задача с двумя классами $Y = \{+1, -1\}$. Рассмотрим линейную модель алгоритмов

$$a(x, w) = \operatorname{sign}\langle x, w \rangle,$$

где $x \in X$ — объект выборки, w — вектор весов признаков.

Рассмотрим специальный случай байесовской классификации, когда предполагается, что признаки имеют вероятностные распределения из экспоненциального семейства. Это дает возможность рассматривать задачи с разнородными исходными

данными, гарантирует линейность NB и аналитическое решение задачи оптимизации параметров распределения по выборке. В этом случае для обучения классификатора достаточно вычислить среднее значение каждого признака в каждом классе. Таким образом, вычислительная сложность алгоритма оптимизации параметров w NB линейна по объему обучающей выборки и по числу признаков. Применение L_1 - регуляризатора (LASSO regression) для отбора признаков улучшает качество классификации, при этом сложность алгоритма обучения остается линейной. L_1 - регуляризатор сводится к методу отбора признаков top-K. Следовательно, NB подходит для анализа больших данных.

Обозначим среднее значение j признака в классе y как

$$\langle x_i^j \rangle_y = \frac{1}{|X_y|} \sum_{x_i \in X_y} x_i^j.$$

Рассмотрим формулы весов признаков:

$$\frac{\langle x_i^j \rangle_{+1}}{\langle x_i^j \rangle_{-1}}; \tag{3.4}$$

$$\ln \frac{\langle x_i^j \rangle_{+1}}{\langle x_i^j \rangle_{-1}}; \tag{3.5}$$

$$\sum_{y \in Y} \sqrt{\langle x_i^j \rangle_y}. \tag{3.6}$$

Эксперименты показали, что наилучшее качество классификации по критерию AUC достигается при использовании формулы (3.6) в качестве формулы вычисления весов признаков объектов. Покажем, что формула весовых коэффициентов (3.6) соответствует одномерным распределениям $p(x^j; \theta_y^j)$ из экспоненциального семейства.

Определение 3.1. *Распределение $p(x)$ принадлежит экспоненциальному семейству распределений [15], если его плотность может быть представлена как:*

$$p(x|\theta, \varphi) = \exp\Big(\frac{x\theta - c(\theta)}{a(\varphi)} + h(x, \varphi)\Big), \tag{3.7}$$

где $c(\theta)$, $h(x, \varphi)$, $a(\varphi)$ — функциональные параметры распределения, θ и φ — числовые параметры, θ — называется параметром сдвига, φ — параметром разброса.

Многие известные распределения относятся к экспоненциальному семейству, в частности:

- нормальное (гауссовское) распределение: $x \in \mathbb{R}$;
- полиномиальное (мультиномиальное) распределение: $x \in \mathbb{N}^+$;
- распределение Бернулли: $x \in \{0, 1\}$;
- биномиальное распределение: $x \in \{0, 1, \ldots, N\}$, где $N \geqslant 0$ — число «испытаний»;
- гамма-распределение: $x \in \mathbb{R}^+$;
- пуассоновское распределение: $x \in \mathbb{N}^+$;
- распределение Лапласа: $x \in \mathbb{R}^+$;
- распределение Дирихле: $x^1, \ldots, x^k$, где $x^j \in (0, 1)$ и $\sum_{j=1}^k x^j = 1$;
- и многие другие.

Пример 3.1. Пусть некоторый признак объектов получен из нормального (гауссовского) распределения $X \sim N(\mu, \sigma^2)$, $E(X) = \mu$, $D(X) = \sigma^2$. Тогда

$$p(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big) = \exp\Big(\frac{\mu x - \frac{\mu^2}{2}}{\sigma^2} + \Big[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}\Big]\Big).$$

Получили, что нормальное распределение принадлежит к экспоненциальному семейству распределений с параметрами $\theta = \mu$, $c(\theta) = \frac{\theta^2}{2}$, $\varphi = \sigma^2$, $a(\varphi) = \varphi$, $h(x, \varphi) = -\frac{1}{2}\log(2\pi\varphi) - \frac{x^2}{2\varphi}$. Нормальным распределением хорошо моделируются действительные признаки объектов $x \in \mathbb{R}$.

Многие методы анализа данных основаны на том, что величины имеют распределения, близкие к нормальному. Нормальное распределение применяют в различных вероятностно-статистических методах принятия решений, например, при статистическом регулировании технологических процессов.

Пример 3.2. Пусть некоторый признак объектов получен из биномиального распределения $X \sim B(n, k)$, где n — число независимых случайных «испытаний», k— вероятность «успеха» в каждом из них. Тогда

$$p(x|n, k) = \binom{n}{x} k^x (1-k)^{n-x} = \exp[x \log \frac{k}{1-k} + n \log(1-k) + \log \binom{n}{x}].$$

Таким образом, биномиальное распределение принадлежит к экспоненциальному семейству распределений с параметрами $\theta = \log \frac{k}{1-k}$, $c(\theta) = n \log(1+e^\theta)$, $\varphi = 1$, $a(\varphi) = 1$,

$h(x,\varphi) = \log \binom{n}{x}$. Биномиальное распределение хорошо подходит для целых неотрицательных признаков объектов с ограниченным числом значений $x \in \{0, 1, \ldots, n\}$.

Биномиальное распределение используется для анализа данных выборочных исследований, в частности, при изучении предпочтений потребителей, выборочном контроле качества продукции по планам одноступенчатого контроля, при испытаниях совокупностей индивидуумов в демографии, социологии, медицине, биологии.

Пример 3.3. Пусть некоторый признак объектов получен из пуассоновского распределения $X \sim Poisson(\lambda)$, $E(X) = \lambda$. Тогда

$$p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} = \exp\big[x\log(\lambda) - \lambda - \log(x!)\big].$$

Следовательно, пуассоновское распределение принадлежит к экспоненциальному семейству распределений с параметрами $\theta = \log(\lambda)$, $c(\theta) = e^\theta$, $\varphi = 1$, $a(\varphi) = 1$, $h(x,\varphi) = -\log(x!)$. Пуассоновское распределение используется, когда признаки объектов $x \in \{0, 1, 2, \ldots\}$.

Распределение Пуассона используется при анализе результатов выборочных маркетинговых обследований потребителей, для описания числа сбоев статистически управляемого технологического процесса в единицу времени, числа «требований на обслуживание», поступающих в единицу времени в систему массового обслуживания, статистических закономерностей несчастных случаев и редких заболеваний.

Пример 3.4. Пусть некоторый признак объектов получен из гамма-распределения $X \sim \Gamma(\alpha, \beta)$, параметр β известен, $E(X) = \frac{\alpha}{\beta}$. Тогда

$$p(x|\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)}e^{-\beta x} = \exp\big[\alpha\log(x) + \alpha\log(\beta) - \log\big(\Gamma(\alpha)\big) - \log(x) - \beta x\big].$$

Предположим параметр α известен. Следовательно, гамма-распределение принадлежит к экспоненциальному семейству распределений с параметрами $\theta = -\beta$, $c(\theta) = -\alpha\log(\beta)$, $\varphi = 1$, $a(\varphi) = 1$, $h(x,\varphi) = \alpha\log(x) - \log\big(\Gamma(\alpha)\big) - \log(x)$. Гамма-распределение используется, когда признаки объектов действительные $x \in [0, +\infty)$.

Во многих случаях гамма-распределению подчинены такие величины, как общий срок службы изделия, длина цепочки токопроводящих пылинок, время достижения изделием предельного состояния при коррозии, время наработки до k-го отказа, $k = 1, 2, \ldots$. Продолжительность жизни больных хроническими заболеваниями, время достижения определенного эффекта при лечении в ряде случаев имеют

гамма-распределение. Это распределение наиболее адекватно для описания спроса в экономико-математических моделях управления запасами.

Ссылаясь на следующие две теоремы, получим, что формула весовых коэффициентов (3.6) соответствует одномерным распределениям из экспоненциального семейства.

Теорема 3.2. *Если одномерные плотности $p(x^j, \theta_j^y)$ принадлежат экспоненциальному семейству распределений и $\Theta = (\theta_j^y)$ является точкой максимума правдоподобия (3.3), то*

$$\theta_y^j = [c']^{-1}(\langle x_i^j \rangle_y). \tag{3.8}$$

Теорема 3.3. *Пусть $Y = \{+1, -1\}$, плотности $p(x^j, \theta_y^j)$ принадлежат экспоненциальному семейству распределений и параметры разброса не зависят от класса, $\varphi_y^j = \varphi^j$. Тогда наивный байесовский классификатор представляется в линейном виде*

$$a(x, w) = \operatorname{sign}\Big(\sum_{j=1}^{n} x^j w_j - w_0\Big),$$

причем веса признаков w_j выражаются через параметры распределений:

$$w_j = \frac{1}{\varphi^j} \sum_{y \in Y} y \theta_y^j, \ j = 1, \ldots, n. \tag{3.9}$$

По формуле (3.9):

$$w^j = \frac{1}{\varphi^j} \sum_{y \in Y} y[c']^{-1}(\langle x_i^j \rangle_y) = \sum_{y \in Y} y \sqrt{\langle x_i^j \rangle_y} = \sqrt{\langle x_i^j \rangle_{+1}} - \sqrt{\langle x_i^j \rangle_{-1}}, \ j = 1, \ldots, n.$$

Отсюда следует, $\varphi^j = \varphi = 1$, $a(\varphi) = 1$, $h(x, \varphi) = 0$, $\theta = \sqrt{\mu}$,

$$\theta = [c']^{-1}(\mu) = \sqrt{\mu}$$

$$c'(\theta) = \mu = \theta^2$$

$$c(\theta) = \frac{\theta^3}{3}.$$

4 Предлагаемый критерий качества алгоритмов выделения терминов

4.1 Стратегии многоклассовой классификации

Многие методы машинного обучения, например, логистическая регрессия, SVM, Байесовские методы, легко обобщаются на случай многих классов. Большинство методов многоклассовой классификации либо базируются на бинарных классификаторах, либо сводятся к ним. Общий подход заключается в использовании набора бинарных классификаторов, обученных разделять различные группы объектов друг от друга. В настоящее время выделяют 3 группы методов многоклассовой классификации [16]:

1. элементарные – простые стратегии, в соответствии с которыми строится набор бинарных классификаторов:
   - один-против-всех;
   - каждый-против-каждого;
   - иерархическая стратегия [17];
2. классические – композиции классификаторов, которые сводятся к стратегии один-против-всех [18] (AdaBoost.MH, AdaBoost.M2, AdaBoost.MR);
3. методы, основанные на кодах с коррекцией ошибки (ECC – error correcting code). К ним относятся: AdaBoost.ECC, AdaBoost.ERP, AdaBoost.ERC.

Для построения многоклассового классификатора в работе использовались стратегии из элементарного подхода и композиция классификаторов из классического. Метод один-против-всех (one-against-all, One-vs-the-rest(OvR)) состоит в построении C классификаторов таким образом, что каждый класс сравнивается с остальными $(C-1)$ классами. Пусть $f(x, w_y) : X \to \mathbb{R}$ — оценка принадлежности объекта $x \in X$ классу y, w_y — весовые коэффициенты признаков. Так как критерий построен на основе линейного классификатора, то

$$f(x, w_y) = \sum_{j=1}^{K} w_y^j x^j. \tag{4.1}$$

Согласно стратегии один-против-всех многоклассовый классификатор записывается:

$$a(x, w) = \arg\max_{y \in Y} f(x, w_y). \tag{4.2}$$

Подход каждый-против-каждого (one-against-one) заключается в построении $\frac{C(C-1)}{2}$ классификаторов таким образом, что каждый класс сопоставляется с каждым из оставшихся. Обозначим оценку принадлежности объекта $x \in X$ классу y против класса z $f(x, w_{yz}) : X \to \mathbb{R}$, заметим, что оценка симметрична $f(x, w_{yz}) = -f(x, w_{zy})$. Итоговый алгоритм стратегии «каждый против каждого»:

$$a(x, w) = \arg\max_{y \in Y} \sum_{z=1}^{|Y|} f(x, w_{yz}). \tag{4.3}$$

В иерархической стратегии классификации обучается бинарное ориентированное дерево с C листьями. Вначале алгоритма все объекты принадлежат одному кластеру (корень). На каждом шаге алгоритма кластеры, содержащие объекты более двух классов, разбиваются на 2 кластера методом кластеризации k-means. Обход графа осуществляется поиском в ширину. Так как объекты одного класса могут попасть в разные кластеры, то необходимо уточнить кластеры K_j:

$$K_j = \{i \in Y : j = \arg\max\{p_{i1}, p_{i2}\}\},\ j = 1, 2, \tag{4.4}$$

где p_{ij} — доля объектов класса i в кластере j.

В каждой вершине дерева настраивается бинарный классификатор. В работе использовался линейный наивный байесовский классификатор. При таком подходе классификатор обучается отличать близкие классы, попавшие в один кластер, поэтому сложнее точно классифицировать объекты, а значит критерий становится более чувствителен к способу представления данных.

4.2 Калибровка классификатора

Калибровка классификатора дает возможность улучшить качество классификации без существенного усложнения процесса вычислений. Кроме того, калибровка позволяет согласовать решения отдельных бинарных классификаторов, которые настраиваются независимо согласно стратегии каждый-против-каждого. Точные, хорошо откалиброванные оценки вероятностей того, что объект x принадлежит классу y, можно интерпретировать как уровень доверия прогнозу. Для получения вероятностей на выходах бинарных классификаторов в исследовании к NB был добавлен

линейный SVM. Заметим, что классификатор по-прежнему остается линейным. Так как SVM имеет тенденцию предсказывать чаще вероятность в середине диапазона, чем на краях, то для получения более точных оценок его необходимо откалибровать.

Существуют три популярных метода калибровки бинарных классификаторов [19]:

- логистическая регрессия (Logistic Regression);
- шкалирование по Платту (Platt Scaling) используется для преобразования результатов SVM от $(-\infty; +\infty)$ к апостериорным вероятностям;
- изотонная регрессия (Isotonic Regression, IR) применяется для усиления «слабых» классификаторов: NB, SVM, решающие деревья.

Для калибровки классификаторов использовался метод, дающий меньшее число ошибок классификации. Эксперименты показали, что лучшим методом по этому критерию оказалась Isotonic Regression. В изотонной регрессии предполагается, что:

$$y_i = m(x_i) + \varepsilon_i,$$

где m — изотонная (монотонно возрастающая) функция.

Задача заключается в том, чтобы по заданной обучающей выборке $X^l = (x_i, y_i)_{i=1}^l$ получить оценку $\hat{m}$ изотонной функции методом наименьших квадратов:

$$\hat{m} = \arg\min_z \sum_{i=1}^{l} (y_i - z(x_i))^2.$$

Isotonic Regression строит кусочно-постоянную аппроксимацию $\hat{m}$. Это можно сделать за линейное время по pair-adjacent violators (PAV) алгоритму [20]. Применение Isotonic Regression на практике для калибровки линейного SVM привело к повышению качества классификации (NB $AUC = 0{,}74$, *NB+SVM+IR* $AUC = 0{,}88$).

4.3 Композиция классификаторов

При решении задачи классификации частей авторефератов применение одного NB с отбором признаков не дает достаточно хорошего качества обучения алгоритма (макро-усреднение (macro average) $AUC = 0{,}66$). Однако, NB прекрасно подошел для классификации целых авторефератов как точная, простая и хорошо интерпретируемая модель. Тогда было решено использовать компромисс — алгоритмическую

композицию классификаторов, которая позволяет разбить исходную задачу на серию подзадач меньшего размера, обучить и затем объединить отдельные базовые алгоритмы NB, тем самым скомпенсировать их прогнозы. Наиболее общее определение алгоритмической композиции дается в алгебраическом подходе к построению корректных алгоритмов, предложенном академиком РАН Ю. И. Журавлёвым [21, 22].

Определение 4.1. *Алгоритмической композицией, составленной из алгоритмических операторов* $b_t : X \to R,\ t = 1, \ldots, T$, *корректирующей операции* $F : R^T \to R$ *и решающего правила* $C : R \to Y$ *называется алгоритм* $a : X \to Y$ *вида*

$$a(x) = C\big(F(b_1(x), \ldots, b_T(x))\big),\ x \in X. \tag{4.5}$$

При построении алгоритмической композиции важно, чтобы базовые алгоритмы как можно сильнее отличались друг от друга. Для усиления различий используются методы:

- метод случайных подпространств (random subspace method, RSM);
- баггинг (bagging — bootstrap aggregation);
- бустинг (boosting).

Учитывая специфику задачи, в работе применялся RSM, так как этот метод:

- лучше для коротких обучающих выборок;
- лучше, когда признаков больше, чем объектов или когда много неинформативных признаков;
- способен повысить точность классификации негибких моделей, не устойчивых к изменениям в признаковом описании;
- подходит для систем, работающих с большими данными (big data), так как обучение ансамблей независимых моделей обладает естественной параллельностью.

RSM параллельно строит равноправные базовые алгоритмы и объединяет их выходы посредством взвешенного голосования [23].

Определение 4.2. *В задаче классификации взвешенным голосованием (weighted voting) называется корректирующая операция F вида*

$$b(x) = F\big(b_1(x), \ldots, b_T(x)\big) = \sum_{t=1}^{T} \alpha^t b_t(x),\ x \in X,\ \alpha^t \in \mathbb{R}.$$

В задаче классификации на M классов $Y = \{1, \ldots, M\}$ решающее правило C относит объект x к тому классу, для которого оценка максимальна:

$$C(b) \equiv C\big(b_1(x), \ldots, b_M(x)\big) = \arg\max_{y \in Y} b_y(x),$$

где пространство оценок $R = \mathbb{R}^M$, алгоритмический оператор $b : X \to \mathbb{R}^M$.

Разнообразие моделей ансамбля в RSM (баггинге) достигается за счет обучения на различных подмножествах признаков (объектов), обычно, случайных. Опишем предложенный алгоритм откалиброванной композиции классификаторов (см. рис. 1). Пусть множество мультиграмм, состоящих из n слов, образуют модальность, где $n = 1, \ldots, 7$, тогда пространство признаков разбивается на отдельные модальности. Для каждой модальности обучается свой многоклассовый классификатор, в работе NB с отбором признаков при стратегии каждый-против-каждого. Затем выходы независимых моделей (значения дискриминантных функций $f(x, w_y)$, $x \in X^k$, X^k — контрольная выборка, $y \in Y$) используются в качестве признаков для линейного SVM, после чего происходит калибровка линейного SVM с помощью IR. Далее необходимо перегруппировать откалиброванные вероятности так, чтобы у каждого объекта было n признаков, соответствующие модальностям, по каждому классу. Это нужно для того, чтобы настроить параметры взвешенного голосования $\alpha^t \in \mathbb{R}$, которые получаются автоматически как веса SVM.

Затем обучается NB на всем пространстве признаков по случайной подвыборке объектов из обучающей выборки. Полученные веса признаков корректируются:

$$w_y^j = \alpha_y^t w_y^j,$$

где $\alpha_y^t \in R$ — параметры взвешенного голосования, $t \in T$, $y \in Y$.

Веса модальностей α_y^t в классе y показывают насколько важны те или иные модальности t для прогнозирования объектов класса y. Некоторые веса оказались равными нулю или даже отрицательными, что говорит о том, что есть модальности признаков, которые не влияют или мешают точной классификации.

Далее выполняется отбор признаков по контрольной выборке. Качество конструкции будет зависеть от качества составляющих ее базовых алгоритмов.

Теоретические оценки достижимой точности основаны на делении квадратичной ошибки на смещение (bias) и вариацию (variance). RSM уменьшает вариацию (дисперсию) и не влияет на смещение композиции. Данная техника равномерно улучшает точность по всем модальностям.

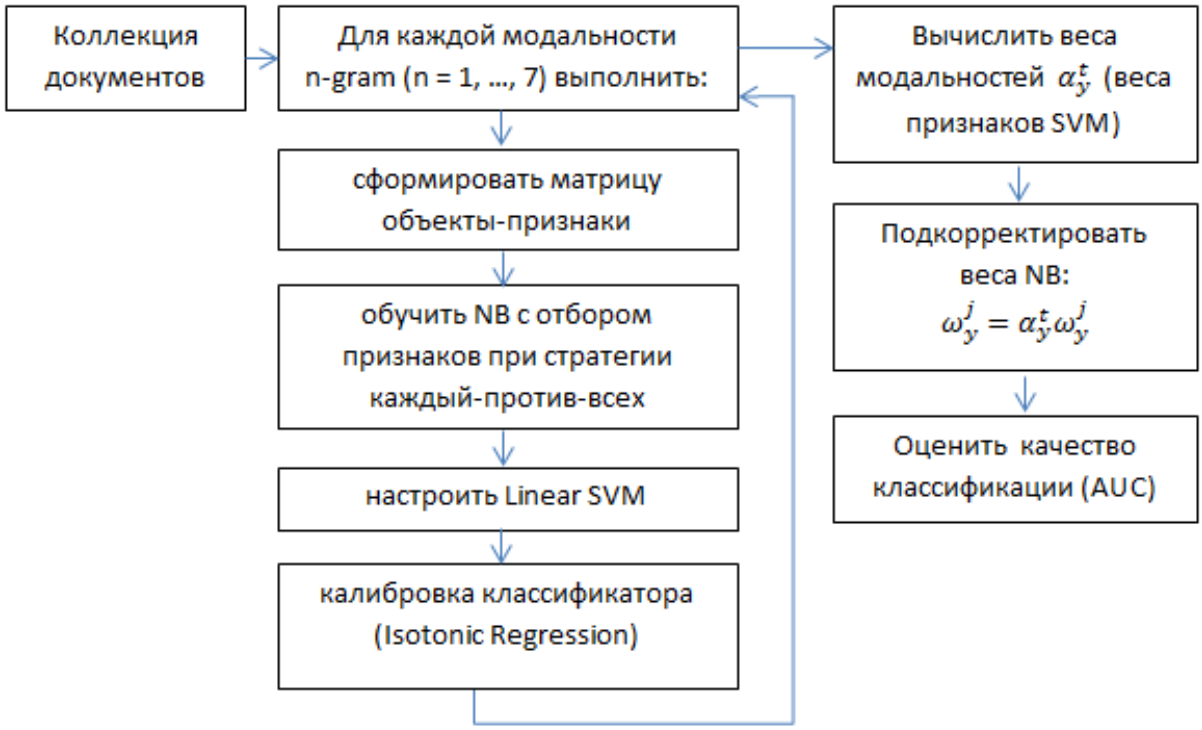


Рис. 1: Алгоритм.

4.4 Описание алгоритма

Исходные документы хранятся в виде «мешка слов»(униграммы или n-граммы). Операцией объединения слов публикаций получаем словарь классификации. С помощью словаря по формуле (2.1) документы преобразуются в бинарные признаки. В результате векторизации документов получим матрицу «объекты-признаки»: строки — объекты (документы публикаций), столбцы — признаки.

Поясним псевдокод алгоритма 4.1. По обучающей выборке X^l вычисляем матрицу весов признаков: вычислить веса w_y^j для всех признаков $j = 1 \dots n$ для каждого класса $y \in Y$. Сортируем признаки по убыванию модуля весов w_y^j.

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 моде-

Алгоритм 4.1. Построение критерия качества алгоритмов *Term Extraction* на основе *NB* за $O(ln)$.

Вход: выборка $X^L = (x_i, y_i)_{i=1}^L$, $X = \mathbb{R}^n$, $Y = \{1, \ldots, C\}$;

Выход: $auc^* = \max_K auc(X^k; K)$,

где X^k — контрольная выборка, параметр $K = (k_1, \ldots, k_m)$, k_i — число информативных признаков i алгоритма классификации;

1: разбить выборку X^L: части документов обучающей выборки целых авторефератов с нечетными номерами относятся к обучающей выборке X^l, с четными — к валидационной выборке X^v, а части документов контрольной выборки целых авторефератов с нечетными номерами относятся к тестовой выборке X^k;

2: обучить веса признаков (пусть для определенности $Y = \{+1, -1\}$)

$$w_y^j = \begin{cases} \sqrt{\langle x_i^j \rangle_{+1}} - \sqrt{\langle x_i^j \rangle_{-1}} & , w_y^j > 0 \\ 0 & , \text{в противном случае} \end{cases}$$

для всех признаков $j = 1 \ldots n$, для каждого класса $y \in Y$;

3: отсортировать признаки по убыванию модуля весов w_y^j, сохранить исходные номера признаков n_y^j;

4: инициализировать $SCORE(X^v, w_y^0) := 0$ для всех $y \in Y$;

5: **для** $j := 1, \ldots, numFeature$

6: $\quad SCORE(X^v, w_y^j) := \sum_{x \in X^v} w_y^j x^{n_y^j} + SCORE(X^v, w_y^{j-1})$ для всех $y \in Y$;

7: $\quad auc_y^j(X^v) := AUC(X^v, SCORE)$ для всех $y \in Y$;

8: инициализировать $k_y := 1$, $auc_y^* := 0$ для всех $y \in Y$;

9: **для** $y := 1, \ldots, C$

10: $\quad k_y := \max_j inverse(auc_y^j)$;

11: $\quad auc_y^* := auc_y^{k_y}(X^k)$;

лей бинарной классификации. По валидационной выборке, используя обученные веса и ответы, вычисляем критерий качества классификации AUC и параметр k_i для каждой модели $i \in \{1, \ldots, m\}$, при котором AUC достигает максимума. Вначале положим $k_i = 1$. Вычисляем для каждой модели дискриминантную функцию ($SCORE$) и критерий качества классификации AUC [24] по алгоритму 2.1. Тестовая выборка X^k, по которой оценивается качество построенной модели, используется для получения несмещенной оцени качества классификации (AUC). Здесь используются обученные

веса признаков и параметр K.

5 Тематическая модель классификации

Стандартные алгоритмы классификации показывают неудовлетворительные результаты на больших текстовых коллекциях с большим числом несбалансированных, взаимозависимых классов [10]. Несбалансированность означает, что классы могут содержать как очень малое, так и очень большое число документов. Взаимозависимые классы имеют схожие множества характерных терминов, и при классификации документа вступают в конкуренцию. Тематические модели лучше справляются с такими задачами, поскольку они учитывают все классы одновременно.

Тематическая модель коллекции текстовых документов определяет множество тем T, к каким темам относится каждый документ и какие слова (термины) образуют каждую тему. Вероятностная порождающая модель выражает вероятности $p(w|d)$ появления терминов w в документах d через распределения $p(w|t)$ и $p(t|d)$. При построение тематической модели требуется оценить по известной коллекции D параметры модели $\varphi_{wt} = p(w|t)$ и $\theta_{td} = p(t|d)$.

Тематическая модель появления слов в документах:

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

где w — слово, t — тема, d — документ коллекции.

Тематическая модель классификации документов:

$$p(c|d) = \sum_{t \in T} p(c|t)p(t|d) = \sum_{t \in T} \psi_{ct}\theta_{td},$$

где c — класс, $c \in C$.

Решается задача максимизации логарифма мультимодального регуляризованного правдоподобия [25]:

$$\sum_{m,d,w} n_{dw} \ln \sum_{t} \varphi_{wt}\theta_{td} + \tau \sum_{d,c} m_{dc} \ln \sum_{t} \psi_{ct}\theta_{td} + R(\Phi, \Theta, \Psi) \to \max_{\Phi,\Theta,\Psi},$$

где $w \in W^m$, W^m — словарь терминов модальности m, $m \in M$, $\Phi = \|\varphi_{wt}\|_{W \times T}$, $\Theta = \|\theta_{td}\|_{T \times D}$, $\Psi = \|\psi_{ct}\|_{C \times T}$, коэффициент регуляризации τ необходим для «приведения к одному масштабу» частот слов n_{dw} и частот классов m_{dc}.

Задача решается с помощью EM-алгоритма, путем модификации E-шага и M-шага [10].

При построении тематической модели использовался подход аддитивной регуляризации тематической модели (ARTM):

$$R(\Phi, \Theta, \Psi) = \sum_i \tau_i R_i(\Phi, \Theta, \Psi).$$

Экспериментально была настроена следующая траектория регуляризации:

- разреживание Θ:
$$R(\Theta) = -\tau \sum_{d \in D} \sum_{t \in T} \ln \theta_{td} \to \max;$$
- сглаживание Φ:
$$R(\Phi) = \tau \sum_{t \in T} \sum_{w \in W} \ln \varphi_{wt} \to \max;$$
- декорреляция тем как столбцов Φ:
$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws} \to \max,$$
где S — предметные темы, $S \subset T$.

Инициализация тематической модели классификации словарем признаков, полученным в результате работы композиции NB, позволяет существенно повысить качество классификации [26] (macro average $AUC = 0{,}99$).

6 Вычислительные эксперименты

6.1 Описание данных

Выборка — коллекция из 8500 авторефератов диссертаций, которые представляют собой тексты на русском языке. Проводились две серии экспериментов, отличающиеся способом нарезания текстов, в обоих случаях в коллекции находились авторефераты с шапкой. В первой серии исследуются целые тексты авторефератов. Половина публикаций была отнесена к обучающей выборке, другая половина — к тестовой. Во-второй серии каждый документ нарезался на 25 равных фрагментов. Разрезалась обучающая выборка целых авторефератов: половина фрагментов была

отнесена к обучающей выборке, другая половина — к валидационной, для контроля использовалась половина частей тестовой выборки. Далее будем уточнять, когда речь идет о частях публикаций.

В задаче 22 класса документов. 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 авторефератов.

Отрасль наук	**Обучающая выборка**		**Тестовая выборка**	
	кол-во авторефератов	доля в выборке, %	кол-во авторефератов	доля в выборке, %
архитектура	4	0,162	4	0,094
биологические науки	172	6,966	305	7,197
ветеринарные науки	13	0,0,527	30	0,708
географические науки	15	0,608	31	0,731
геолого-минералогические науки	26	1,053	48	1,133
искусствоведение	18	0,729	33	0,779
исторические науки	116	4,698	195	4,601
культурология	9	0,365	19	0,448
медицинские науки	523	21,183	885	20,882
педагогические науки	127	5,144	242	5,710
политические науки	27	1,094	59	1,392
психологические науки	30	1,215	54	1,274
сельскохозяйственные науки	67	2,714	124	2,926
социологические науки	42	1,701	61	1,439
технические науки	353	14,297	613	14,464
фармацевтические науки	7	0,284	16	0,378
физико-математические науки	198	8,019	321	7,574
филологические науки	145	5,873	226	5,333
философские науки	82	3,321	148	3,492
химические науки	66	2,973	111	2,619
экономические науки	350	14,176	577	13,615
юридические науки	79	3,200	136	3,209

В данной работе сравниваются два вида признаков: униграммы и мультиграммы. На вход алгоритма подаются словари коллекций текстов, один из них состоит из 139229 униграмм, другой содержит 667566 мультиграмм. Униграммы были получены

после следующей обработки:

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

Для автоматического выделения мультиграмм были рассмотрены три алгоритма Term Extraction. В *Termhood* [5] применялась тематическая модель к задаче извлечения терминов (мультиграмм) коллекций текстовых документов. Алгоритм состоит из двух этапов. На первом этапе формируется избыточно большой лексикон из мультиграмм, отобранных по морфологическим признакам и статистическим критериям релевантности и устойчивости мультиграмм. На втором этапе отбираются мультиграммы, наиболее полезные для тематической модели, что позволяет существенно сократить лексикон без ухудшения качества тематической модели.

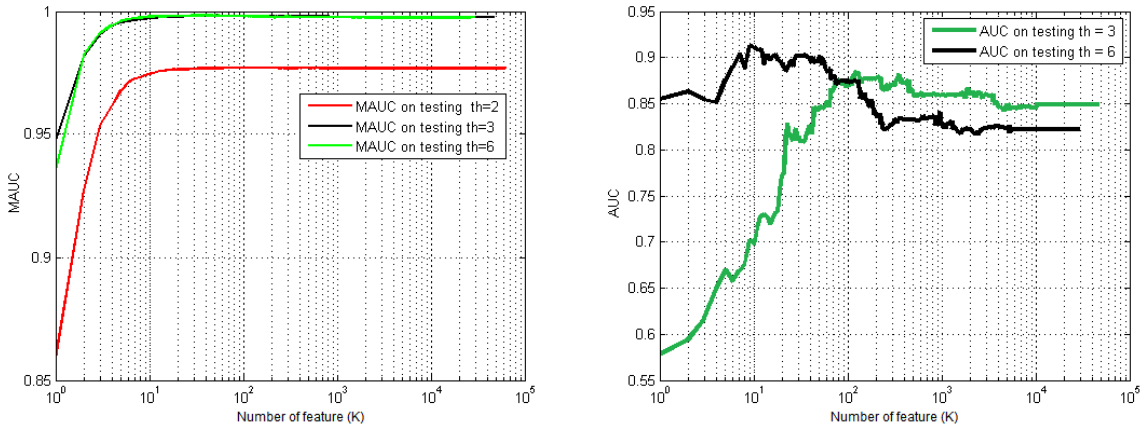
6.2 Выбор порога встречаемости терминов

Цель эксперимента: подобрать оптимальный, с точки зрения максимизации критериев $AUC/MAUC$, порог встречаемости терминов в авторефератах th.

Порог подбирался посредством классификации униграмм (см. рис. 2):

- при стратегии каждый-против-каждого использовался критерий $MAUC$;
- по критерию AUC для классификатора пары классов культурология и философские науки, имеющего самое низкое качество классификации.

Эксперименты показали, что оптимальным порогом встречаемости термина в документе является три, так как качество классификации при $th = 3$ и $th = 6$ практически одинаковое, но при пороге 6 уменьшается обобщающая способность классификатора. При классификации частей использовался порог 2 или 3.



(a) MAUC в зависимости от числа признаков. (b) Культурология против философские науки.

Рис. 2: Выбор порога встречаемости *th*.

6.3 Оценивание качества выделения терминов авторефератов

Цель эксперимента: применяя NB с различными стратегиями многоклассовой классификации, сравнить униграммную языковую модель с мультиграммной.

В ходе экспериментов были рассмотрены униграммная и мультиграммная модели признакового описания объектов. В качестве стратегии разбиения на классы использовались стратегии:

- каждый-против-всех;
- каждый-против-каждого;
- иерархическая стратегия.

Вычислялся AUC на контрольной выборке (при иерархической стратегии — на обучающей) в зависимости от числа признаков K.

На многих графиках AUC 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 контроле лучше, чем на обучении.

Эксперименты показали, что использование мультиграмм в качестве признаков наивного байесовского классификатора приводит к улучшению качества классификации (см. рис.5 – рис.9). Сравнение построенного классификатора NB с использованием метода top-K с линейным SVM приводится в таблице (см. рис.10). При представлении многоклассового классификатора как совокупность бинарных классификаторов возникает необходимость вычислить агрегированный показатель качества классификации, объединяющий показатели отдельных классификаторов. Для этого существует два метода. При макроусреднении (macro average) вычисляется взвешенное среднее значение по классам. При микроусреднении (micro average) объединяются решения на уровне документов по всем классам.

Кроме того, разработанный критерий качества применялся для оценивания и сравнения различных алгоритмов Term Extraction. Критерий, построенный на основе качества классификации, позволяет отбирать термины предметных областей для каждого алгоритма Term Extraction. В условиях эксперимента величина критерия качества классификации AUC объективно отражает способность алгоритмов Term Extraction выделять мультиграммы. Отметим, что длина фрагмента текста не влияет на качество выполняемых измерений. Проведенные нами эксперименты показали, что *Termhood* [5] выделяет мультиграммы лучше по макро-усреднению (macro average) и микро-усреднению (micro average) в обеих сериях экспериментов (на документах полностью и фрагментах).

6.3.1 Стратегия каждый-против-каждого

Цель эксперимента: исследовать с помощью NB насколько хорошо отличаются классы друг от друга.

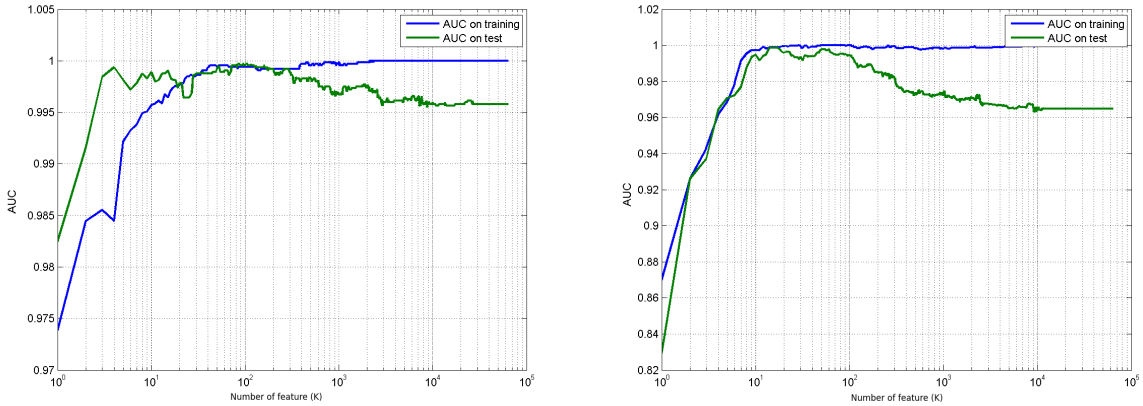
Большинство классов хорошо отличаются друг от друга (см. рис. 4). Наихудшим качеством классификации обладал классификатор культурология против философские науки, но его удалось улучшить, применяя мультиграммы. Обозначим для определенности $Y = \{+1, -1\}$. На графиках, изображенных на рис. 3, для обучения весов признаков применялась формула (6.1). Однако, использование таких весов в качестве

критерия отбора признаков дает в топе не только термины предметной области, но и термины другого класса, имеющие большие отрицательные веса в нашем классе $+1$, то есть эти термины противопоставляют наш класс $+1$ другому -1 (другим в случае стратегии каждый-против-всех). Цель состоит в том, чтобы отбирать термины областей наук, поэтому в дальнейшем применялся как критерий отбора признаков (6.2), который приводит к незначительному ухудшению качества классификации, зато менее ччувствителен к составу сравниваемых классов.

$$w_y^j = \sqrt{\langle x_i^j \rangle_{+1}} - \sqrt{\langle x_i^j \rangle_{-1}} \tag{6.1}$$

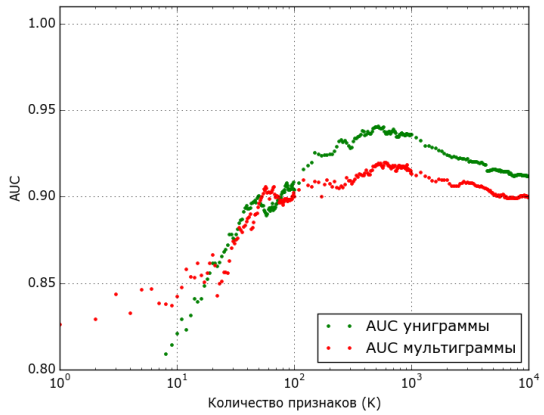
$$w_y^j = \begin{cases} \sqrt{\langle x_i^j \rangle_{+1}} - \sqrt{\langle x_i^j \rangle_{-1}} & , w_y^j > 0 \\ 0 & , \text{в противном случае} \end{cases} \tag{6.2}$$

для всех признаков $j = 1 \ldots n$.

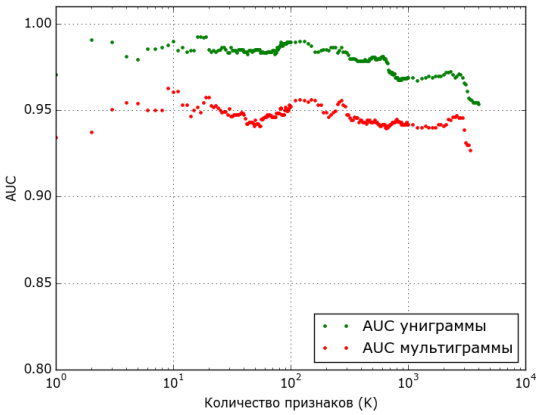


(a) Искусствоведение против исторических наук. (b) Политические науки против социологических.

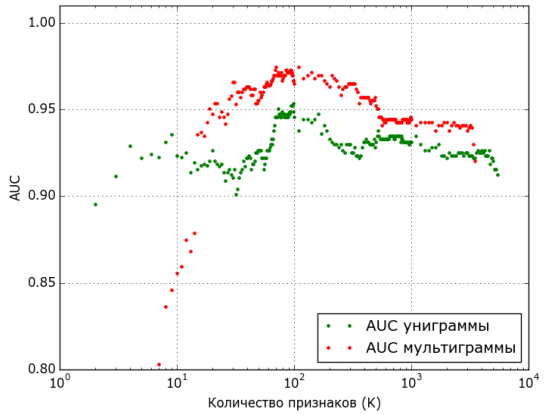
Рис. 3: Переобучение отбора признаков в униграммной модели.



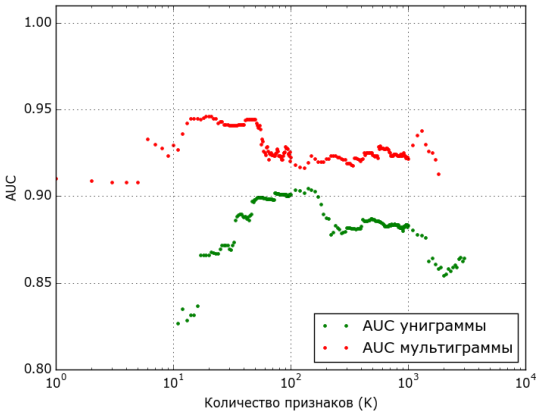
(a) Биологические науки против ветеринарных.



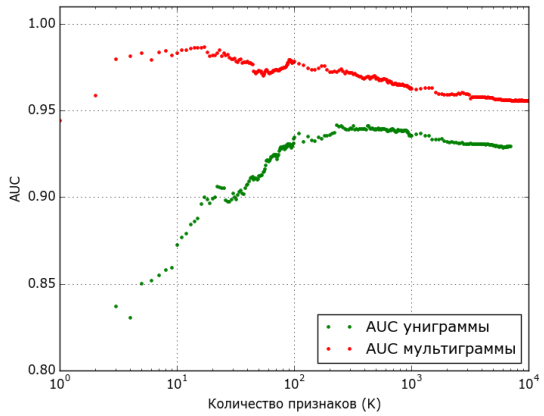
(b) Географические науки против геолого- минералогических.



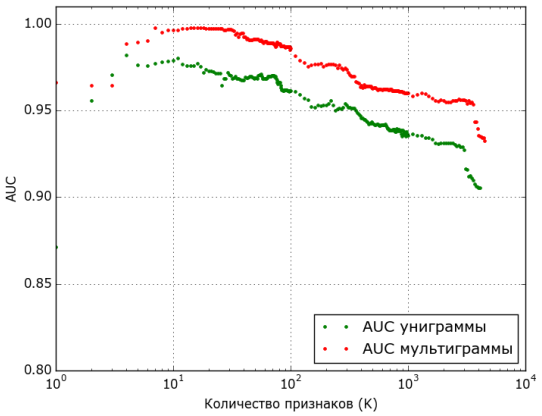
(c) Искусствоведение против культурологии.



(d) Культурология против философских наук.



(e) Педагогические науки против психологических.



(f) Политические науки против социологических.

Рис. 4: Сравнение униграммной и n-граммной моделей при использовании стратегии каждый-против-каждого.

6.3.2 Стратегия каждый-против-всех

Цель эксперимента:

1. исследовать с помощью NB насколько хорошо отличается каждый класс от всех остальных класов;
2. изучить отобранные информативные признаки (термины) предметных областей.

При стратегии каждый-против-всех всего методом top-K было отобрано 223 информативных мультиграммы, из них уникальных — 200. Значения параметра $K = \{k_1, \ldots, k_{22}\}$ находятся в диапазоне от 5 до 48 признаков для различных классов.

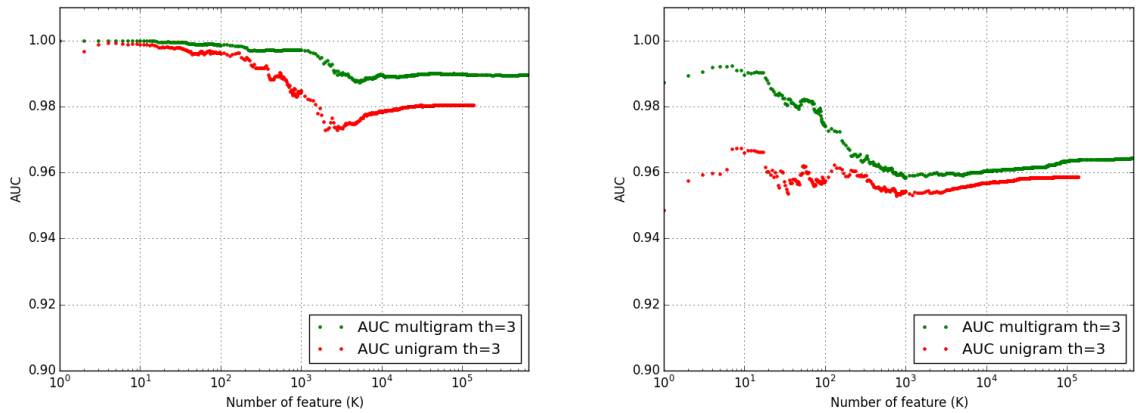
Приведем список отобранных признаков (первые 5) для некоторых классов:

- класс биологические науки:
  1. биологический наука 0.89
  2. биологический 0.62
  3. диссертационный исследование -0.54
  4. идея -0.52
  5. биология 0.51
- класс исторические науки:
  1. доктор исторический наука 1.00
  2. исторический наука 0.93
  3. историография 0.82
  4. хронологический 0.76
  5. историк 0.71
- класс филологические науки:
  1. наука филологический 0.96
  2. филологический 0.90
  3. филология 0.69
  4. слово 0.67

5. языковой 0.65

- класс юридические науки:

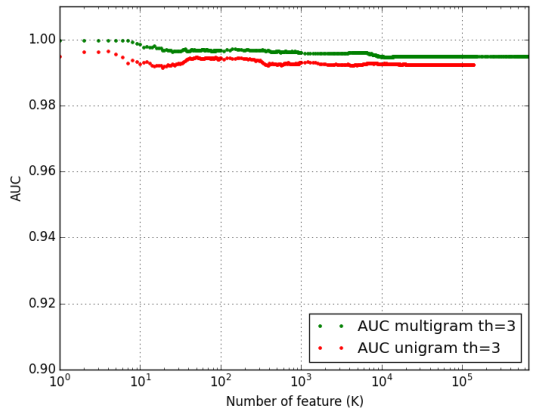
1. наука юридический 0.95
2. правовой регулирование 0.78
3. акт правовой 0.78
4. юридический 0.76
5. законодатель 0.76



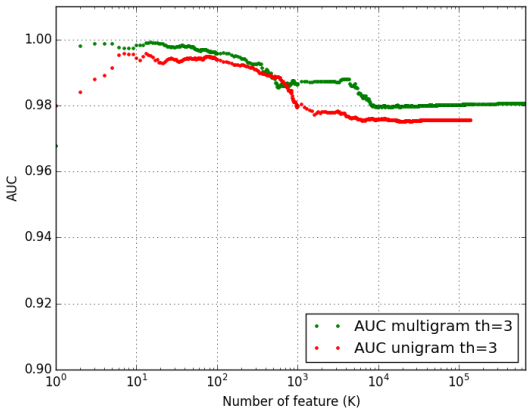
(a) Класс архитектура.

(b) Класс биологические науки.

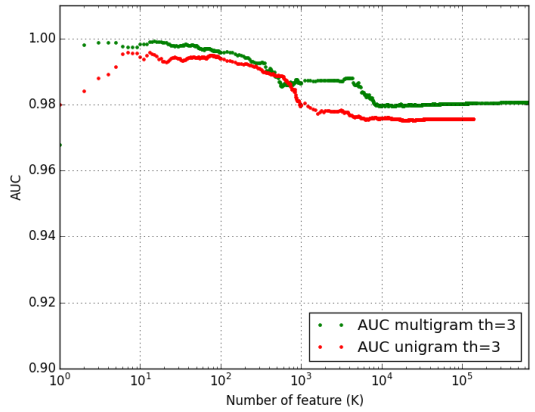
Рис. 5: Сравнение униграммной и n-граммной моделей при использовании стратегии каждый-против-каждого.



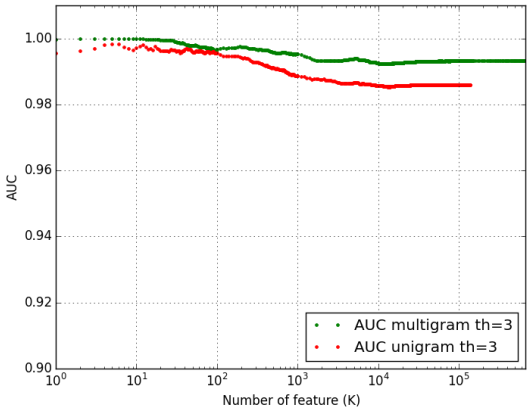
(a) Класс ветеринарные науки.



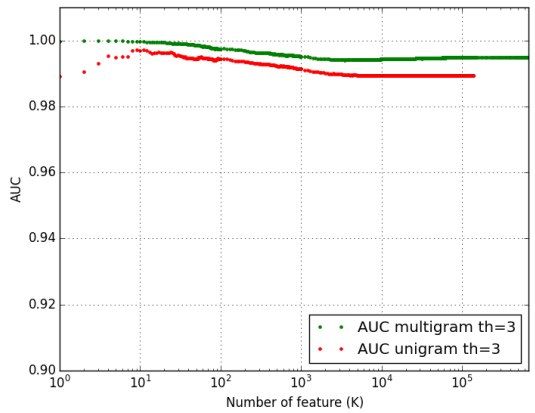
(b) Класс географические науки.



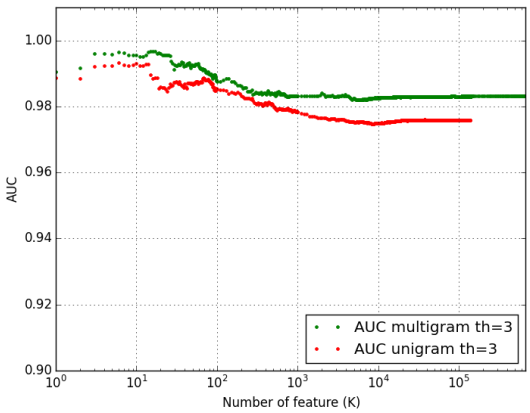
(c) Класс геолого-минералогические науки.



(d) Класс искусствоведение.

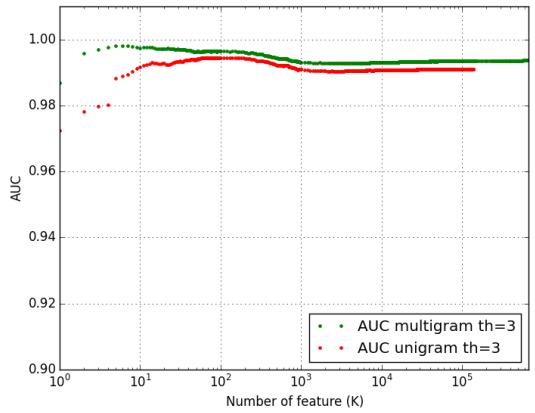


(e) Класс исторические науки.

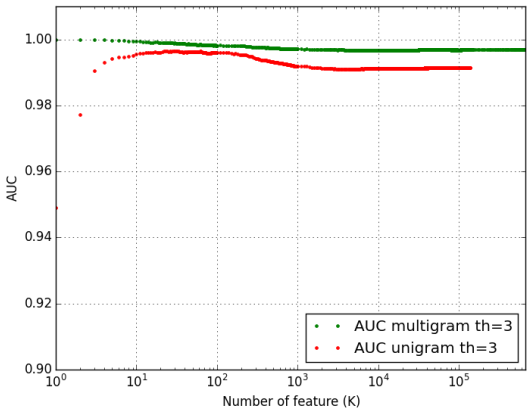


(f) Класс культурология.

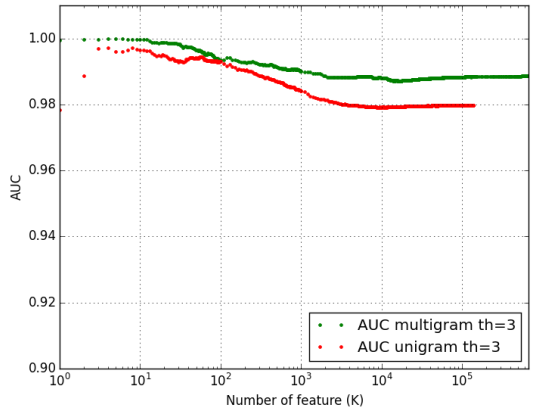
Рис. 6: Сравнение униграммной и n-граммной моделей при использовании стратегии один-против-всех.



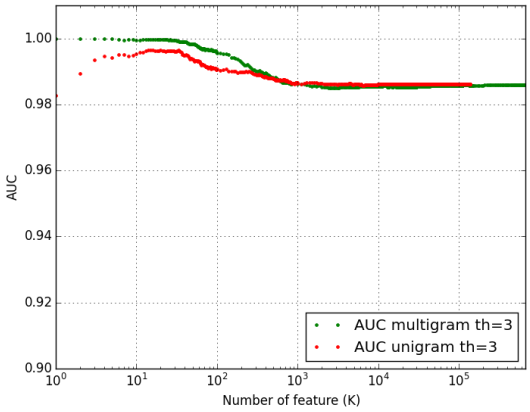
(a) Класс медицинские науки .



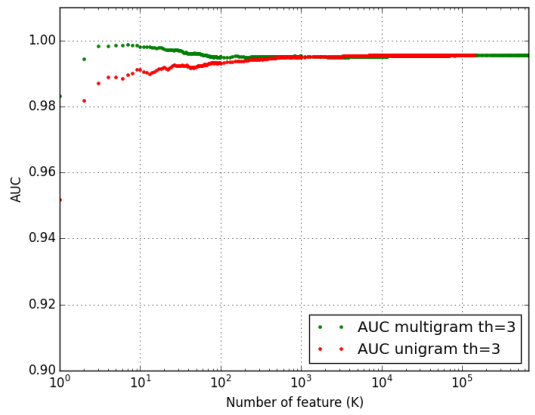
(b) Класс педагогические науки.



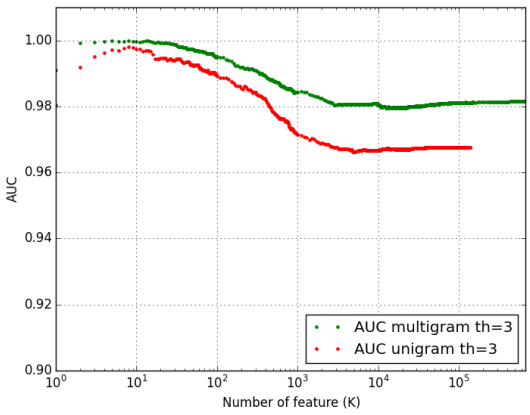
(c) Класс политические науки.



(d) Класс психологические науки.

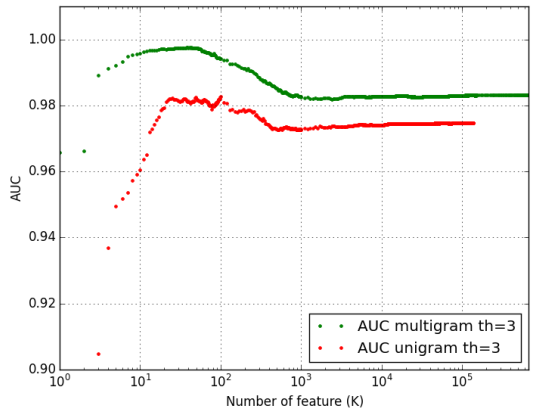


(e) Класс сельскохозяйственные науки.

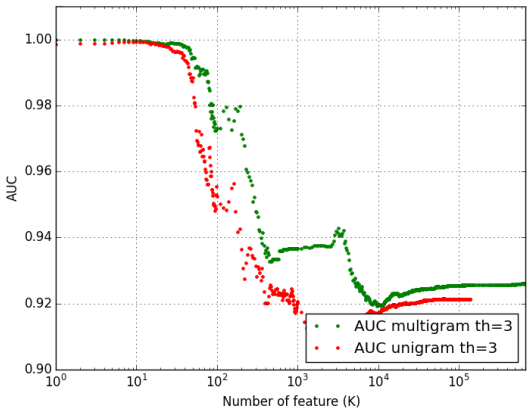


(f) Класс социологические науки.

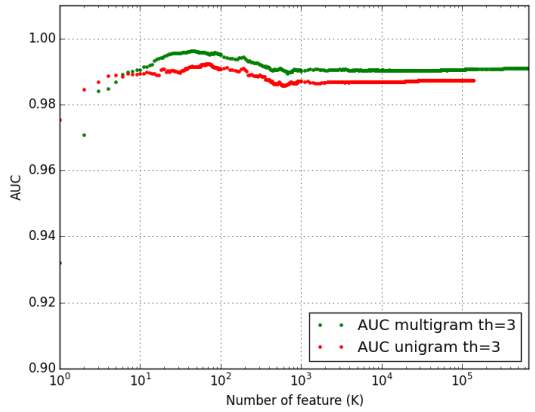
Рис. 7: Сравнение униграммной и n-граммной моделей при использовании стратегии каждый-против-каждого.



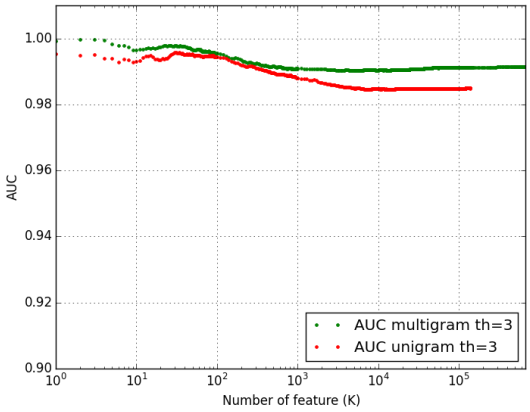
(a) Класс технические науки.



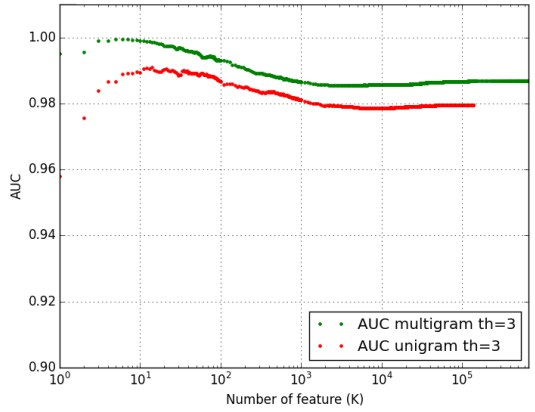
(b) Класс фармацевтические науки.



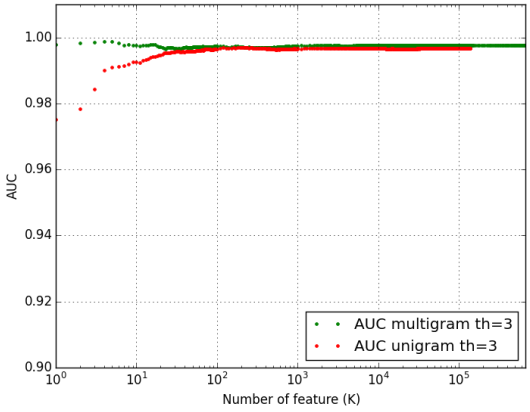
(c) Класс физико-математические науки.



(d) Класс филологические науки.

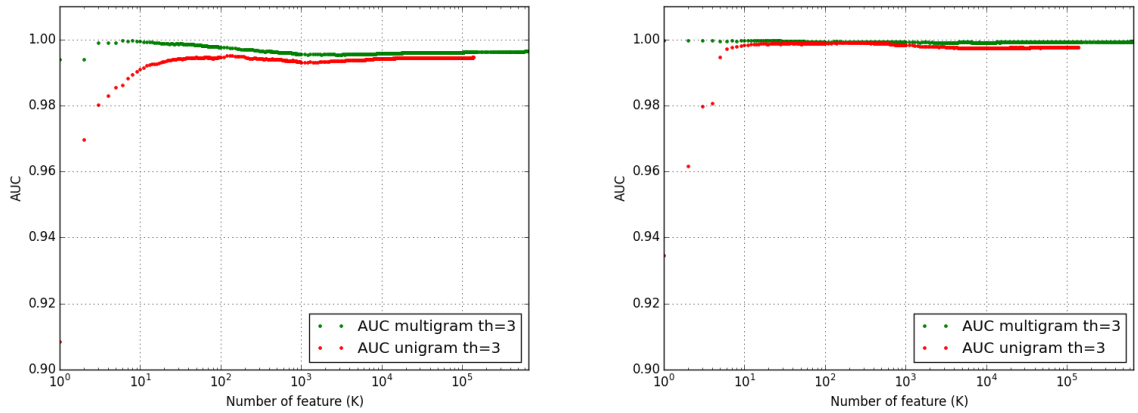


(e) Класс философские науки.



(f) Класс химические науки.

Рис. 8: Сравнение униграммной и n-граммной моделей при использовании стратегии каждый-против-каждого.



(a) Класс экономические науки.

(b) Класс юридические науки.

Рис. 9: Сравнение униграммной и n-граммной моделей при использовании стратегии каждый-против-каждого.

Классы	n-gram		unigram	
	NB_topK	SVM	NB_topK	SVM
архитектура	1,0000	0,9991	0,9993	0,9984
биологические науки	0,9923	0,9907	0,9584	0,9791
ветеринарные науки	0,9998	0,9984	0,9972	0,9980
географические науки	0,9993	0,9958	0,9914	0,9946
геолого-минералогические науки	0,9999	0,9996	0,9988	0,9991
искусствоведение	1,0000	0,9998	1,0000	0,9998
исторические науки	0,9999	0,9998	0,9953	0,9995
культурология	0,9967	0,9907	0,9848	0,9769
медицинские науки	0,9980	0,9982	0,9938	0,9962
педагогические науки	0,9999	0,9995	0,9947	0,9990
политические науки	0,9999	0,9976	0,9939	0,9953
психологические науки	1,0000	0,9994	0,9955	0,9992
сельскохозяйственные науки	0,9987	0,9979	0,9926	0,9960
социологические науки	0,9999	0,9993	0,9973	0,9993
технические науки	0,9975	0,9988	0,9844	0,9961
фармацевтические науки	1,0000	0,9987	0,9708	0,9976
физико-математические науки	0,9963	0,9988	0,9920	0,9985
филологические науки	0,9994	0,9997	0,9969	0,9997
философские науки	0,9994	0,9987	0,9898	0,9957
химические науки	0,9988	0,9988	0,9943	0,9977
экономические науки	0,9996	0,9997	0,9951	0,9988
юридические науки	0,9998	0,9999	0,9984	0,9998
macro average	0,9989	0,9981	0,9916	0,9961
micro average	0,9937	0,9985	0,9668	0,9972

Рис. 10: Сравнение NB с линейным SVM по униграммным и n-граммным признакам.

6.3.3 Иерархическая стратегия

Цель эксперимента:

1. исследовать с помощью NB насколько хорошо отличаются близкие классы, попавшие в один кластер (вершина дерева);
2. изучить отобранные информативные признаки (термины) предметных областей.

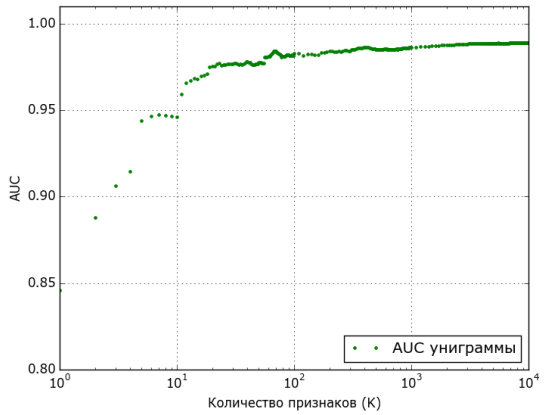
В эксперименте использовалась униграммная языковая модель. При иерархической стратегии становится сложнее различать классы (см. рис. 12), так как противопоставляются близкие классы, попавшие в один кластер. Многоклассовый классификатор при такой стратегии становится более чувствительным к методу выделения n-грамм измерительным инструментом, но имеет более трудоемкую реализацию, поскольку требуется построить бинарное дерево (см. рис. 11). Приведем первые 5 признаков, попавших в список отобранных признаков для некоторых классов:

- естественно-научные науки против гуманитарных наук:
    1. температура 0.55
    2. частота 0.55
    3. препарат 0.52
    4. клинический 0.51
    5. раствор 0.51
- экономические науки против гуманитарных наук:
    1. инвестиционный 0.67
    2. инвестиция 0.64
    3. затрата 0.64
    4. стоимость 0.62
    5. расчет 0.60
- медицинские науки против ветеринарных наук:
    1. пациент 0.88

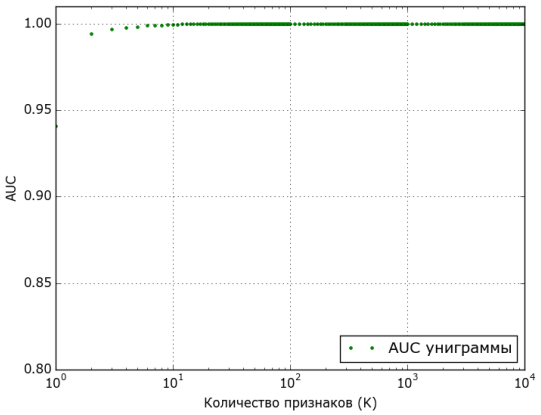
2. здравоохранение 0.69
3. медицинский 0.69
4. женщина 0.68
5. лицо 0.66

- юридические науки против политики, истории, архитектуры:

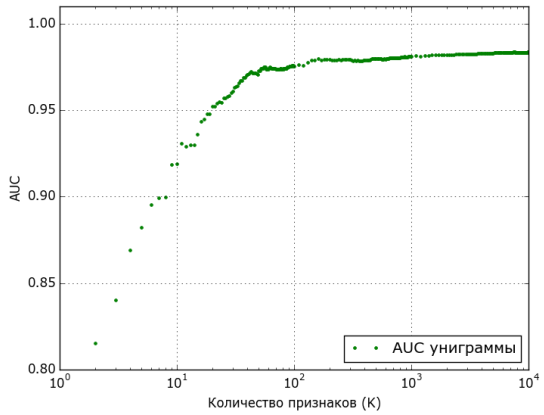
1. судебный 0.77
2. юридический 0.76
3. суд 0.75
4. законодательство 0.74
5. законодатель 0.74



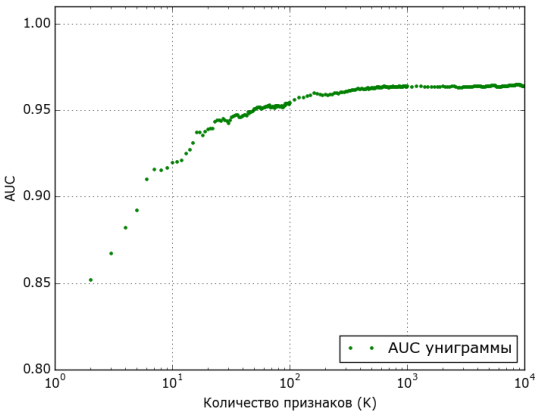
(a) Класс архитектура.



(b) Класс геолого-минералогические науки.



(c) Класс биологические науки.



(d) Класс искусствоведение.

Рис. 12: Качество классификации на обучении при иерархической стратегии.

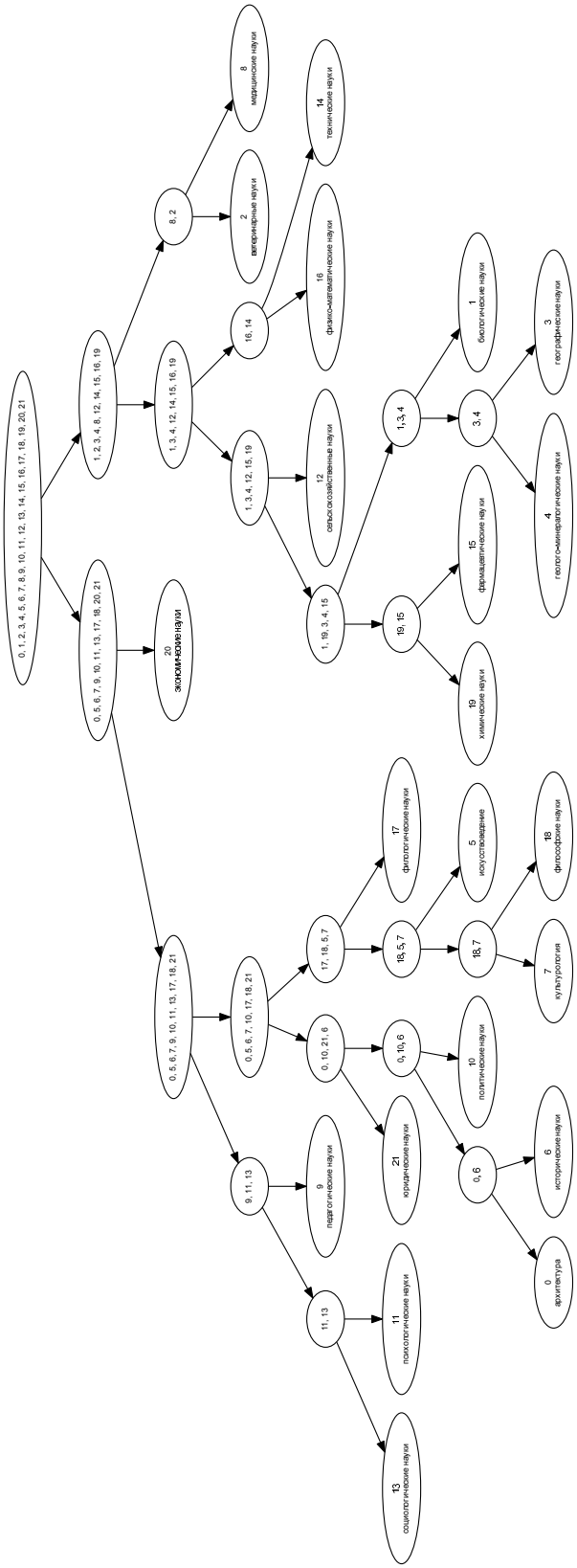


Рис. 11: Обучение многоклассового классификатора при иерархической стратегии.

6.4 Оценивание качества выделения терминов частей авторефератов

Цель эксперимента:

1. усложнить задачу для алгоритмов Term Extraction тем, что термины нужно выделять во фрагментах авторефератов;

2. построить более чувствительный критерий качества алгоритмов Term Extraction на основе качества классификации.

6.4.1 Фильтация по документной частоте

Цель эксперимента: проверить, можно ли использовать фильтрацию признаков по документной частоте для улучшения качества классификации фрагментов авторефератов по критерию AUC.

Эксперимент показал, что использование фильтрации признаков по документной частоте приводит к сильному уменьшению словаря, числа отобранных признаков и как следствие ухудшению качества классификации (см. табл. 2):

1. словарь n-грамм без фильтрации по DF: 253338 признаков, словарь n-грамм с фильтрацией по DF: 46636 признаков;

2. число отобранных признаков без фильтрации по DF в диапазоне $430-21000$, а с фильтрацией по DF диапазон значительно уже $1-4000$ и может быть сильное переобучение (вырожденный случай с 1 признаком).

Для некоторых классов фильтрация по DF привела к увеличению AUC, но какой-либо закономерности между числом объектов в классе и AUC не выявлено.

Таблица 2: Сравнение качества классификации частей авторефератов в зависимости от фильтрации по документной частоте.

Отрасль наук	Композиция классификаторов	
	с фильтрацией по DF	без фильтрации по DF
архитектура	0,8972	0,9311
биологические науки	0,8013	0,7855
ветеринарные науки	0,8923	0,8592
географические науки	0,8058	0,8536
геолого-минералогические науки	0,9087	0,9302
искусствоведение	0,9090	0,9606
исторические науки	0,8986	0,9164
культурология	0,8244	0,9087
медицинские науки	0,8040	0,8668
педагогические науки	0,8367	0,8897
политические науки	0,8791	0,9369
психологические науки	0,7732	0,8436
сельскохозяйственные науки	0,8130	0,8294
социологические науки	0,8250	0,8848
технические науки	0,7857	0,8342
фармацевтические науки	0,8271	0,6869
физико-математические науки	0,8578	0,8535
филологические науки	0,8690	0,9266
философские науки	0,8788	0,9102
химические науки	0,9183	0,9222
экономические науки	0,7845	0,8528
юридические науки	0,9680	0,9592
macro average	0,8526	0,8792
micro average	0,8109	0,8526

Метод	NB	Композиция		
		NB+wt SVM	ANN	NB+IR+wt SVM
macro average AUC	0,749	0,865	0,876	0,887
micro average AUC	0,733	0,873	0,880	0,880

Рис. 13: Сравнение NB с композициями NB.

tau	0,001	0,01	0,1	0,2	0,3	0,4
macro average AUC	0,5848	0,595124	0,645414	0,641802	0,639504	0,630563
micro average AUC	0,641044	0,625293	0,723555	0,728487	0,699332	0,696013

Рис. 14: Оптимизация параметра регуляризации SVM, максималтное число итераций = 700.

6.4.2 Композиция классификаторов

Цель эксперимента: точнее настроить веса NB с помощью композиции NB.

Композиция позволяет точнее настраивать веса NB за счет учета влияния модальности n-грамм (см. рис. 13). Мультиграммы получены на 1 этапе алгоритма. Лучшее качество классификации среди композиций у *откалиброванной композиции NB классификаторов*.

6.4.3 Сравнение моделей классификации

Цель эксперимента:

1. исследовать как качество выделения терминов влияет на качество классификации;

2. выбрать чувствительный критерий качества алгоритмов Term Extraction.

Эксперимент проводился на подвыборке фрагментов авторефератов.

Описание подвыборки:

X^l — обучающая выборка, $|X^l| = 30000$

X^k — контрольная выборка, $|X^k| = 5000$

$x = (x^1, \ldots, x^n)$, $n \geqslant 1$.

В работе оптимизировался параметр регуляризации SVM (см. рис. 14). Реализация SVM была взята из библиотеки Scikit-Learn для Python.

Лучшее качество классификации фрагментов авторефератов показала тематическая модель классификации (см. рис. 15), построенная с помощью подхода ARTM

macro average AUC

Метод	Композиция NB	SVM	ARTM	ARTM (иниц. комп. NB)
1 этап	0,8865	0,927608693	0,9415	0,9914
TF-IDF	0,8382	0,879496049	0,9360	0,9902
Termhood	0,8941	0,908999405	0,9496	0,9987

micro average AUC

Метод	Композиция NB	SVM	ARTM	ARTM (иниц. комп. NB)
1 этап	0,8477	0,946084471	0,9554	0,9898
TF-IDF	0,8412	0,913888121	0,9479	0,9970
Termhood	0,8597	0,932079654	0,9488	0,9990

Рис. 15: Сравнение моделей классификации на подвыборке объектов.

Метод	1 этап	TF-IDF	Termhood
macro average AUC	0,8792	0,8267	0,9018
micro average AUC	0,8526	0,8468	0,8773

Рис. 16: Сравнение качества выделения терминов фрагментов авторефератов.

и проинициализированная признаками, отобранными откалиброванной композицией NB.

Как критерий качества алгоритмов Term Extraction была выбрана композиция NB, так как:

1. качество классификации чувствительно к способу представления данных;

2. требуется меньше вычислительных ресурсов для обучения алгоритма.

Дальнейшее сравнение проводилось с помощью композиции NB, оценивалось качество выделения терминов фрагментов публикаций на всей коллекции (см. рис. 16).

6.5 Зависимость AUC от длины фрагмента

Цель эксперимента: выявить закономерность AUC от длины фрагмента для различных методов выделения терминов.

В распределении длин фрагментов (см. рис. 17) большинство фрагментов содержат менее 200 признаков.

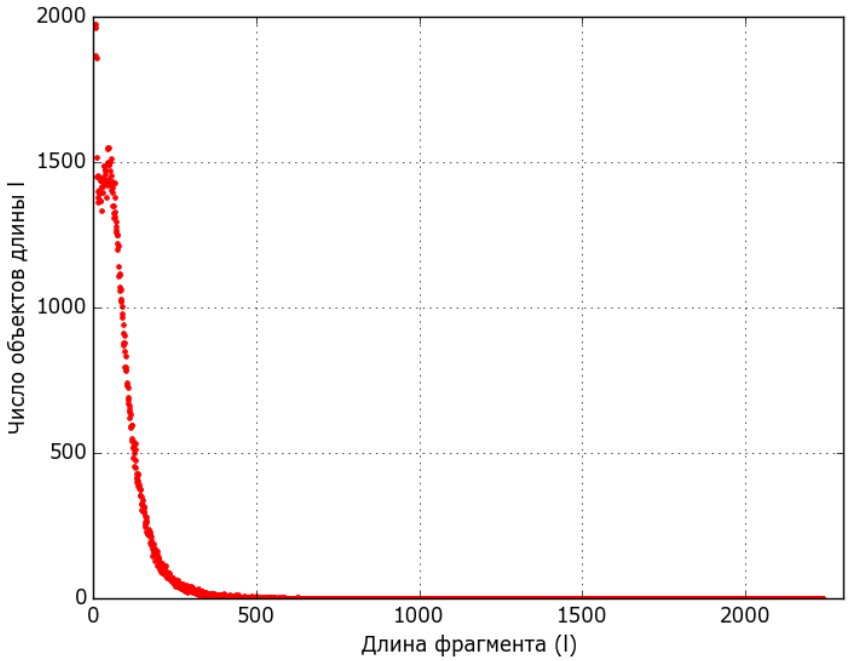


Рис. 17: Распределение длин фрагментов.

На графиках (рис. 18 и 19) можно увидеть, что методы первого этапа лучше работают на фрагментах с числом признаков менее 30, а для всех остальных длин качество классификации на всей коллекции фрагментов лучше у метода Termhood.

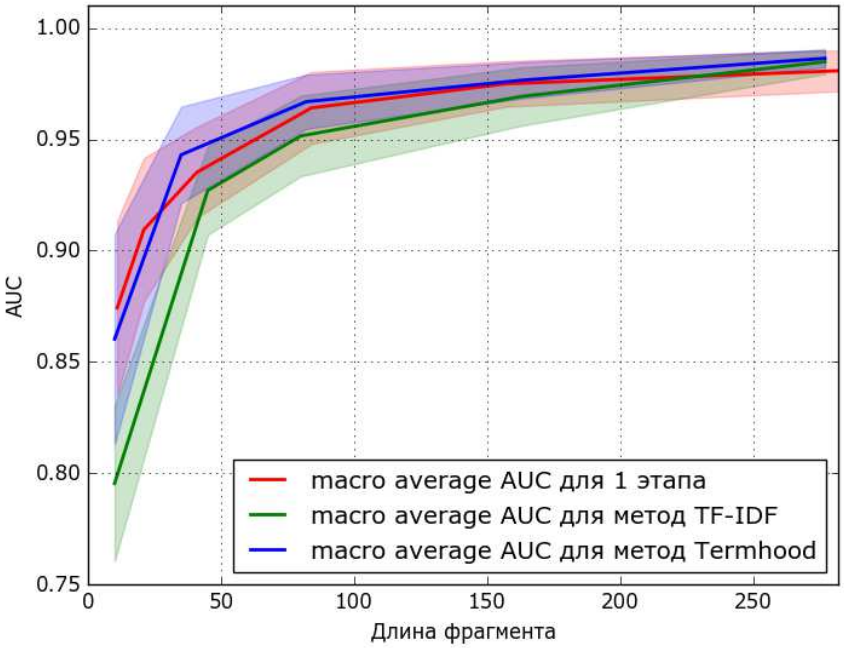


Рис. 18: Зависимость macro average AUC от длины фрагмента.

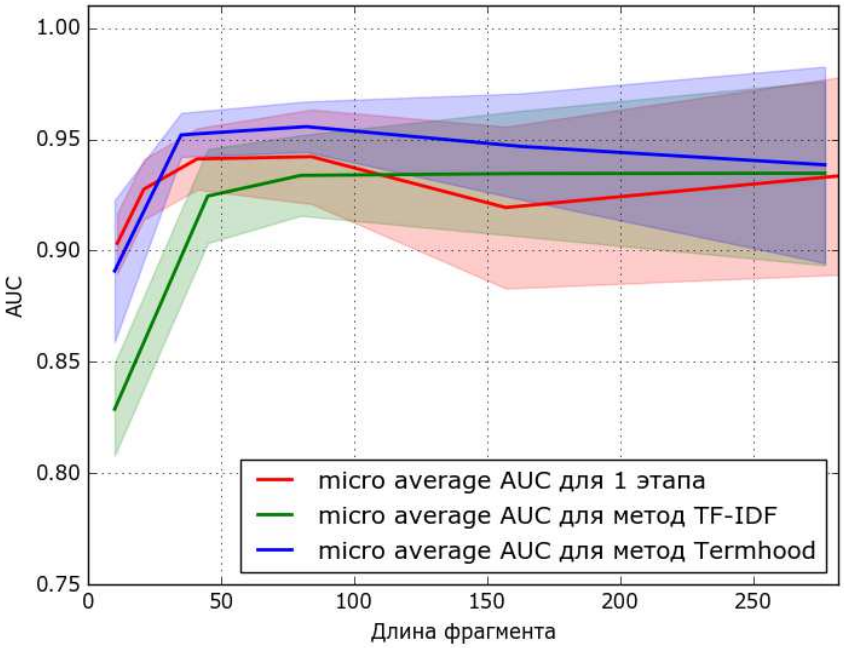


Рис. 19: Зависимость micro average AUC от длины фрагмента.

7 Заключение

В работе предложен критерий для оценивания и сравнения алгоритмов выделения терминов, основанный на качестве классификации композиции наивных байесовских классификаторов. Реализованы алгоритмы классификации: NB, композиция NB, тематическая модель классификации (с помощью библиотеки BigARTM). Композиция NB позволяет быстро и эффективно сравнить алгоритмы выделения терминов. Для дальнейшего повышения качества классификации можно проинициализировать тематическую модель классификации признаками, отобранными композицией NB, и получить AUC равный 0,99.

Проведены вычислительные эксперименты по сравнению униграммной и мультиграммной моделей классификации, которые подтвердили, что n-граммы дают лучшие результаты, чем униграммы, по критерию качества AUC.

Дальнейшие исследования связаны с рассмотрением других методов на втором этапе алгоритма Term Extraction для того, чтобы выбрать из них лучший. Результаты второго этапа будут использоваться на третьем этапе алгоритма.

Список литературы

[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] 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.

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

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

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

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

[10] КВ Воронцов. Вероятностное тематическое моделирование. *Москва*, 2013.

[11] Наталья Валентинова Лукашевич. Тезаурусы в задачах информационного поиска. In *М.: Издательство МГУ, 2011*. М.: Издательство МГУ, 2011, 2010.

[12] Владимир Борисович Барахнин, Дмитрий Александрович Ткачев, et al. Кластеризация текстовых документов на основе составных ключевых термов. *Вестн. НГУ. Сер. Информ. технологии*, 8(2):5–14, 2010.

[13] Tang K. Wang R. An empirical study of mauc in multi-class problems with uncertain cost matrices. *arXiv preprint arXiv:1209.1800*, 2012.

[14] Харт П. Дуда Р. *Распознавание образов и анализ сцен.* М.:Мир, 1976.

[15] Clark D. R. and Thayer C. A. A primer on the exponential family of distributions. *Casualty Actuarial Society Spring Forum*, pages 117–148, 2004.

[16] Баринова О. Соболев А. Многоклассовая классификация в задаче семантической сегментации. In *Graphicon*, 2009.

[17] F. Schwenker. Hierarchical support vector machines for multi-class pattern recognition. *Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, pages 561 – 565 vol.2, 30 Aug 2000-01 Sep 2000 2000.

[18] Robert Tibshirani Jerome Friedman, Trevor Hastie. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, Volume 28(№ 2):337–407, 2000.

[19] Caruana R. Niculescu-Mizil A. Predicting good probabilities with supervised learning. *Proceedings of the 22nd international conference on Machine learning. – ACM*, pages 625–632, 2005.

[20] G. Ewing W. Reid and E M. Ayer, H. Brunk. An empirical distribution function for sampling with incomplete information. *Annals of Mathematical Statistics*, page 641–647, 1955.

[21] Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации. *Проблемы кибернетики*, Т. 33:5–68, 1978.

[22] К. В. Воронцов. *Лекции по алгоритмическим композициям.* 2007 г.

[23] К. В. Воронцов. Комбинаторный подход к оценке качества обучаемых алгоритмов. *Математические вопросы кибернетики*, 13:5–36, 2004.

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

[25] Timothy N Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine learning*, 88(1-2):157–208, 2012.

[26] Evgeny Sokolov and Lev Bogolubsky. Topic models regularization and initialization for regression problems. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, pages 21–27. ACM, 2015.