Классификация научных текстов по отраслям знаний

А. В. Сухарева

Научный руководитель: В. В. Стрижов
Консультант: С. В. Царьков
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

25 мая 2015 г.

Цели проекта

к окончанию курса разработать алгоритм для исследования качества работы различных методов автоматического выделения терминов.

Применение

классификация документов активно применяется в работе электронных рубрикаторов документов, информационных поисковых систем. Качество решения данной задачи сильно влияет на результат работы указанных систем.

Что предлагается

оценить качество работы линейного многоклассового классификатора с отбором признаков (униграмм и n-грамм), имеющим линейное по числу объектов и числу признаков время обучения.

Постановка задачи

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.
Дано:
$X^l = (x_i, y_i)_{i=1}^l$ — обучающая выборка документов.
Ставится задача восстановления зависимости $y = f(x)$ по точкам обучающей выборки X^l.
Предполагается, что документы $x \in X$ описываются бинарными признаками $(x^1, \ldots, x^N)$:

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

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

Постановка задачи

Гипотеза

Для каждого класса важна только небольшая часть признаков K, 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.

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

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

Постановка задачи

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

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

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

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

Постановка задачи

Для выбора других классов будем использовать стратегии каждый-против-каждого.
В работе наивный байесовский классификатор используется для построения линейного многоклассового классификатора:

$$a(x) = \arg\max_{y} \sum_{w=1}^{K} wt(w, y) x^w, x \in X, y \in Y, \tag{5}$$

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.

Постановка задачи

Функционал качества $MAUC$ (мультиклассовый AUC):

$$MAUC = \frac{1}{c}\sum_{i=1}^{c}\frac{1}{c-1}\sum_{j=1,j\neq i}^{c} A_{ij}, \tag{6}$$

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

$$AUC = \frac{1}{l_0 l_1}\sum_{i=1}^{l}\sum_{j=1}^{l}[y_i > y_j][a(x_i,w) < a(x_j,w)], \tag{7}$$

где $l_0 = \sum_{i=1}^{l}[y_i = 0]$, $l_1 = \sum_{i=1}^{l}[y_i = 1]$.

Алгоритм

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 (5).
На выходе алгоритма получаем: AUC, $MAUC$ на обучении и контоле в зависимости от параметра K.
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.
По обучающей выборке X^l вычисляем матрицу весов признаков.

Алгоритм

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*.

Вначале положим $k_i = 1$. Вычисляем для каждой модели критерий качества классификации *AUC*.

Вычисляем критерий *MAUC* по формуле (5) при каждом зачении k_i.

Запоминаем оптимальный k_i.

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*.

Цели вычислительного эксперимента

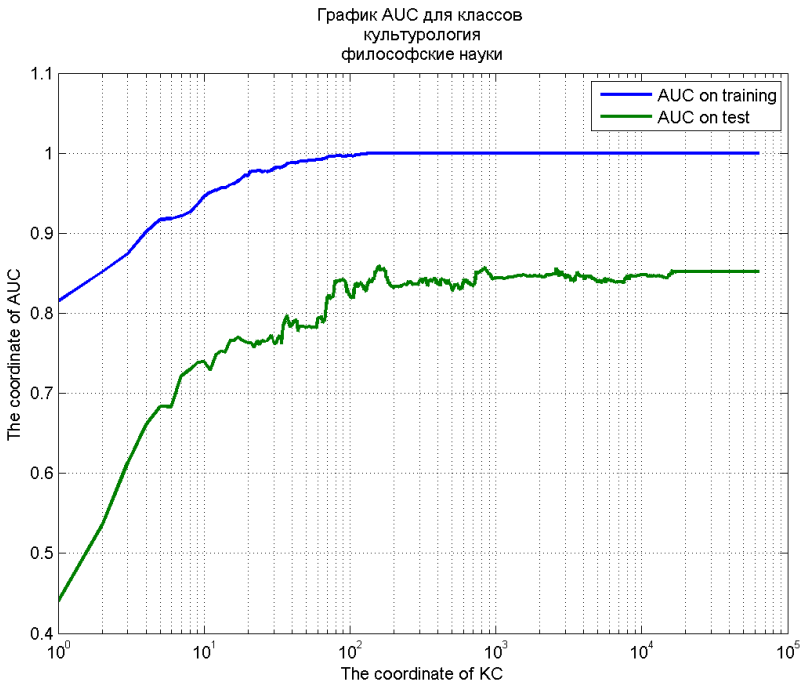
1. Сравнить качество классификации (AUC/MAUC) по каждому классу в зависимости от числа первых K признаков.
2. Отобрать для каждого класса информативные признаки.

Для этого был проведен эксперимент:

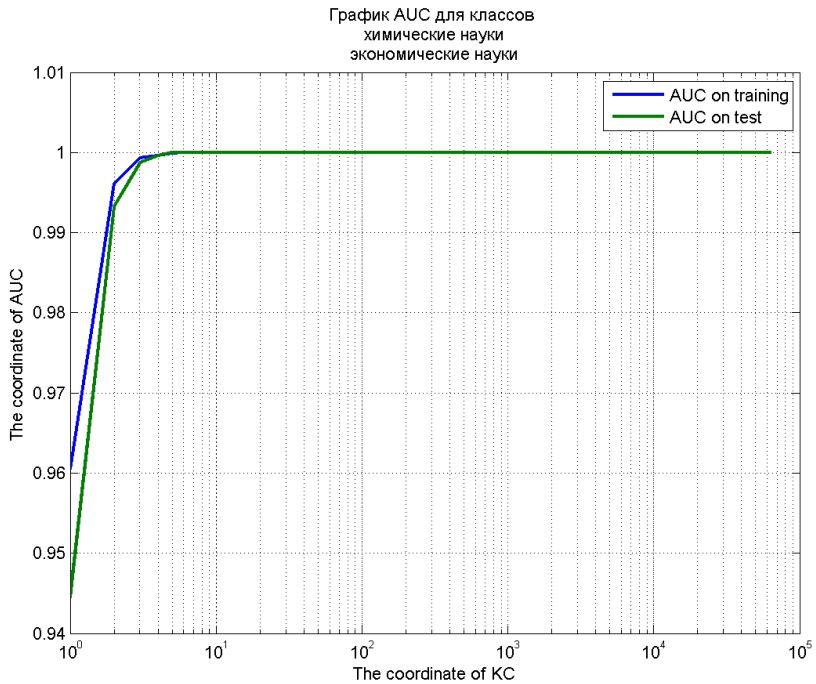
- Выборка — коллекция из 7852 авторефератов диссертаций на русском языке.
- В задаче рассматривается 20 классов документов. Классами являются области науки.
- Данная задача является задачей с несбалансированными классами (unbalanced classes).

Результаты вычислительного эксперимента

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



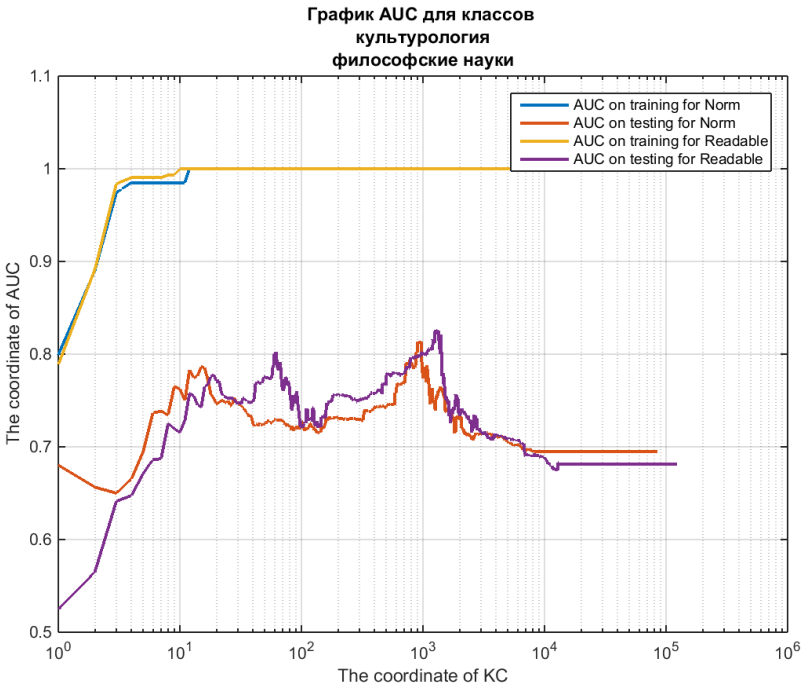
(a) Плохо различимые классы.



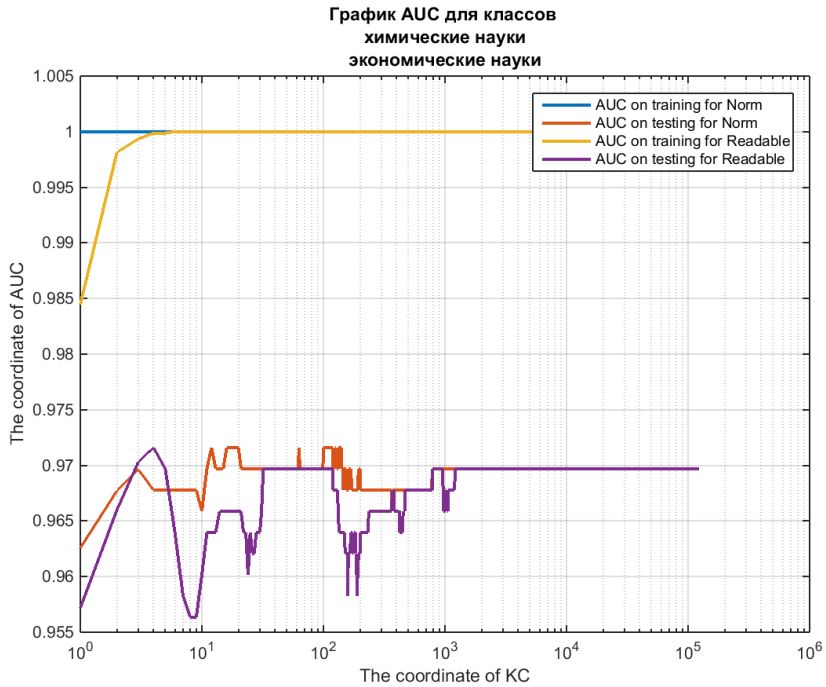
(b) Хорошо различимые классы.

Рис.: Степень различимости классов в униграммной модели.

Результаты вычислительного эксперимента



(a) Плохо различимые классы.

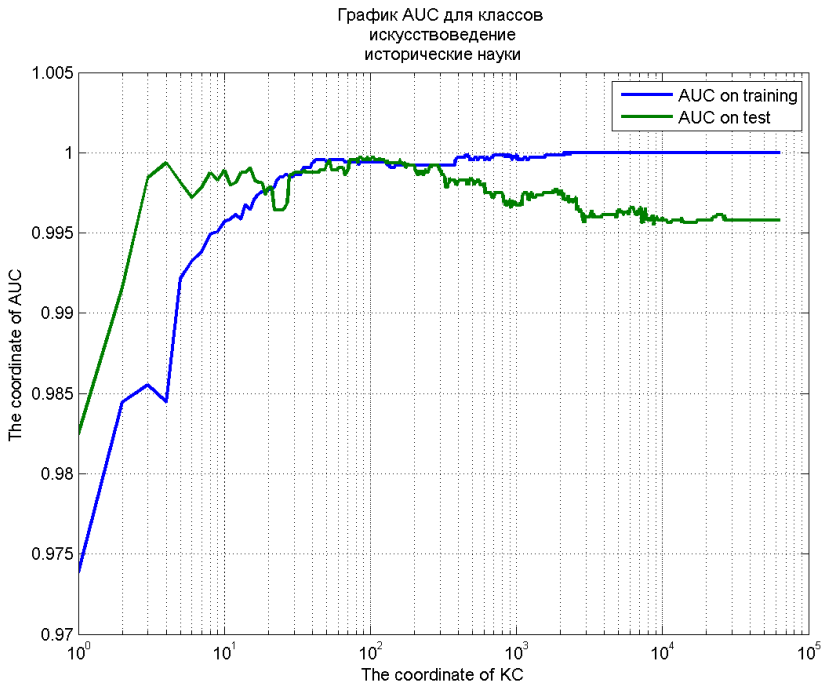


(b) Хорошо различимые классы.

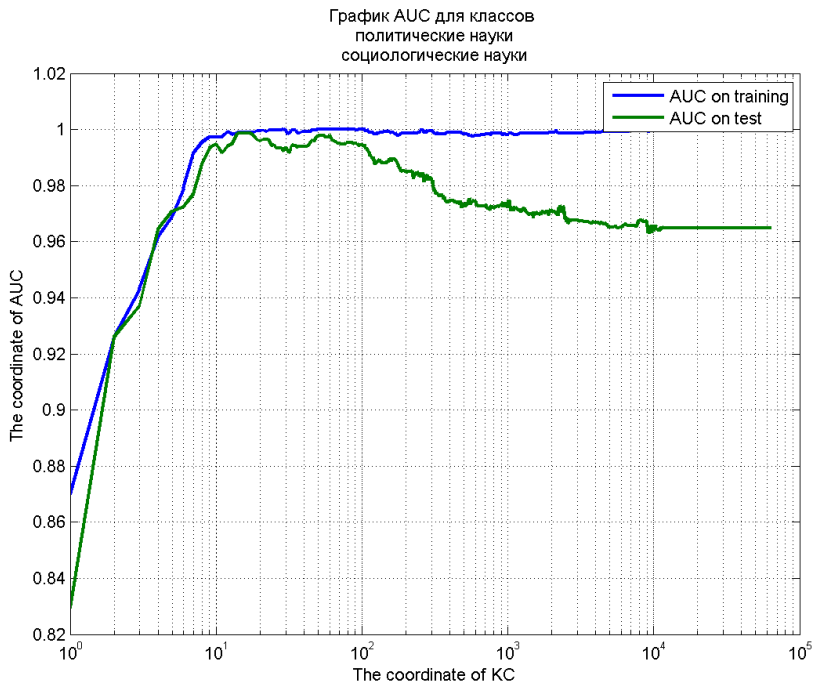
Рис.: Степень различимости классов в n-граммной модели.

Результаты вычислительного эксперимента

Использование n-грамм в качестве признаков наивного байесовского классификатора приводит к ухудшению качества классификации.



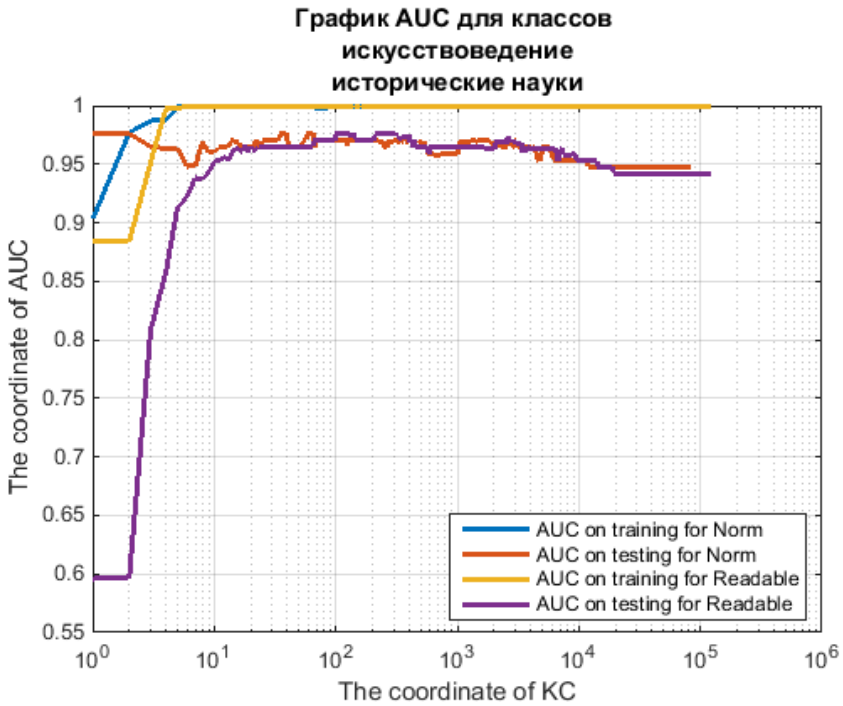
(a) Сильное влияние шума при небольшом числе признаков.



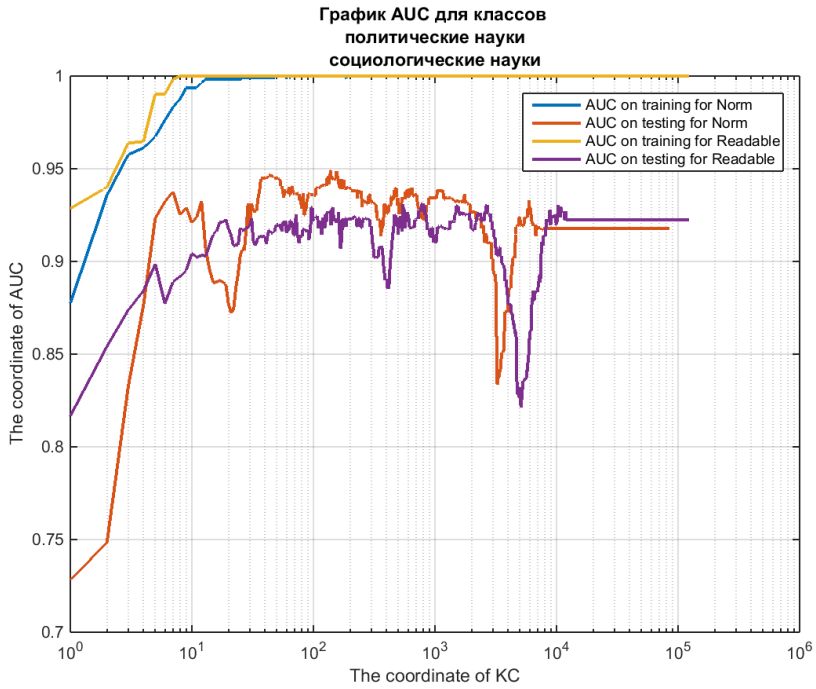
(b) Шум не влияет на критерий качества.

Рис.: Переобучение отбора признаков в униграммной модели.

Результаты вычислительного эксперимента



(a) Сильное влияние шума при небольшом числе признаков.

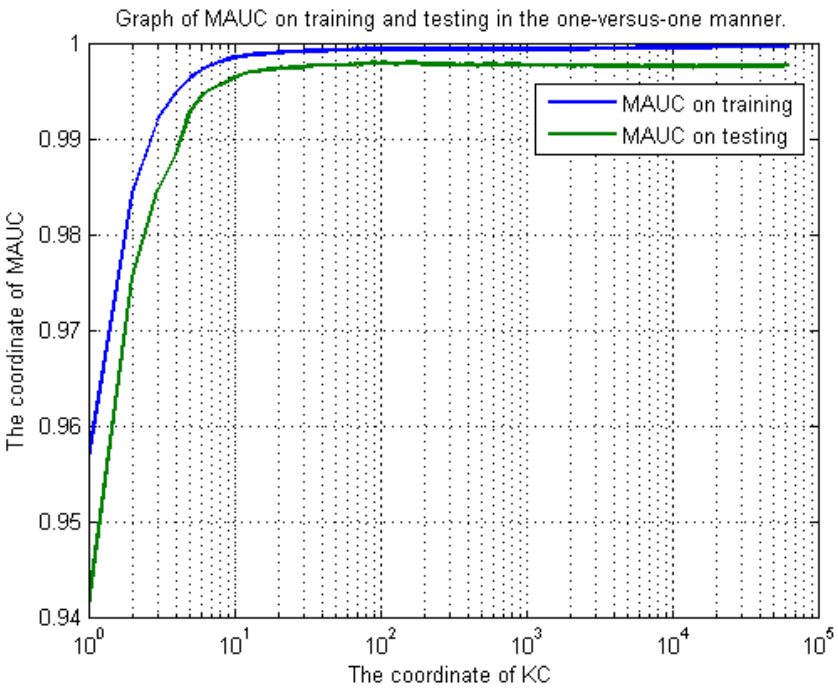


(b) Шум не влияет на критерий качества.

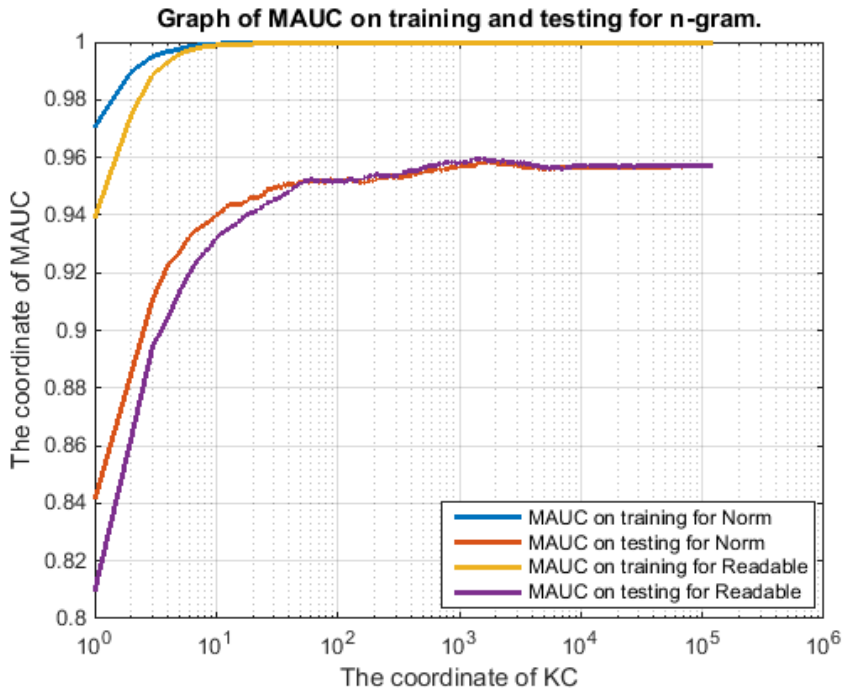
Рис.: Переобучение в n-граммной модели.

Результаты вычислительного эксперимента

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



(a) Униграммная модель.



(b) *n*-граммная модель.

Рис.: Мультиклассовый *AUC*.

Заключение

Эксперименты, проведённые на русскоязычных авторефератах диссертаций, показывают, что униграммы дают лучшие результаты, чем n-граммы, по критерию качества $MAUC$. Модель «мешка слов» имеет ряд недостатков. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов в документах друг от друга.

Добавления словосочетаний, n-грамм и многословных терминов в тематические модели часто приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели.

Комбинация униграмм и биграмм позволяет улучшить результаты.

Литература

1. Курс лекций К.В. Воронцова. Машинное обучение.
2. Лукашевич Н.В., Логачев Ю.М. Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. - Т2 - 2010. - стр. 108-116.
3. Царьков С.В. Автоматическое выделение ключевых фраз для построения словаря терминов в тематических моделях коллекций текстовых документов. Естественные и технические науки №6(62)., стр. 456-464., 2012.