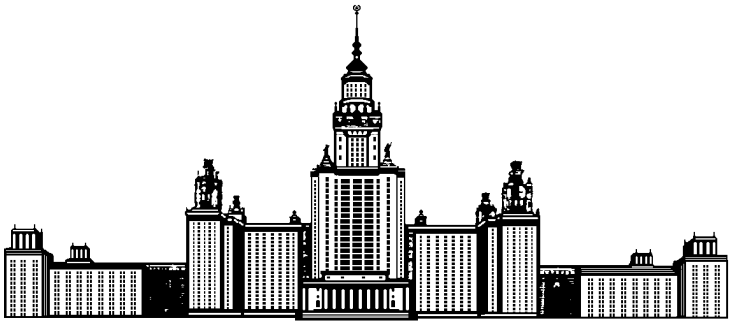
Московский государственный университет имени М.В. Ломоносова
Факультет Вычислительной Математики и Кибернетики



Магистерская программа
«Логические и комбинаторные методы анализа данных»

Магистерская диссертация

«Классификация тем в вероятностных тематических моделях коллекций текстовых документов»

Работу выполнил:
Шапулин Андрей Валентинович

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

Москва
2017

Содержание

Аннотация

Тематические модели являются популярным инструментом анализа текстовой информации. Они позволяют автоматически выделить структуру в огромных коллекциях документов. Однако, не все темы, получаемые при помощи тематического моделирования, являются качественными и интерпретируемыми. Также часто возникает задача поиска определенной тематики в документах. Для того, чтобы найти релевантные темы или измерить качество нахождения предметных тем обычно привлекается человеческий труд асессоров. В данной работе описывается алгоритм построения автоматического классификатора тем, который призван облегчить труд эксперта. Эксперименты проводились на различных тематических моделях с размеченной информацией о связанности тем с международными или межэтническими отношениями.

1 Введение

Тематическая модель является инструментом поиска структуры в текстовых данных, однако, она не предоставляет информации о верхнеуровневом описании тем, которая могла бы быть понятна человеку. Для больших коллекций документов оптимальное число тем может достигать тысяч, и, соответственно, чтобы найти интерпретируемые или связанные с искомой тематикой темы, требуется большое количество человеческих ресурсов.

За время активного использования тематических моделей накопилось достаточно данных о разметке тем. Логичным дальнейшим шагом является построение автоматического классификатора тем, который способен если не исключить человека из процесса понимания тем, то существенно облегчить его труд, предоставляя оценки вероятности принадлежности темы некоторой категории.

В простейшем случае алгоритм может решать бинарную задачу классификации: отделять фоновые темы от предметных, либо искать темы, связанные с конкретной тематикой.

На сегодняшний день существуют сотни вариаций алгоритмов тематических моделей [6]. Поэтому важно построить алгоритм, который может одинаково хорошо классифицировать темы независимо от способа их построения.

Для обучения классификатора в данной работе используются различные модели с размеченными темами, в частности, рассматривается базовая модель латентного размещения Дирихле (LDA) [3], а также модели, построенные с использованием подхода аддитивной регуляризации тематических моделей (далее APTM) [12].

В работе решается задача отделения тем, связанных с межэтническими и международными отношениями от всех остальных при помощи алгоритмов машинного обучения. Данные для обучения представляют собой темы тематические моделей, построенных на коллекции постов популярного портала LiveJournal и статей сервиса IQBuzz. Классы тем размечались асессором вручную.

2 Тематическое моделирование

Вероятностная тематическая модель (probabilistic topic model) коллекции текстовых документов осуществляет одновременную кластеризацию документов и слов по кластерам-темам. При этом кластеризация является «мягкой». Это означает, что каждый документ может относиться к нескольким темам, тогда как обычные методы кластеризации основаны на чрезмерно жёстком предположении монотематичности доку-

ментов. Каждое слово также может относиться к нескольким темам, что позволяет избежать проблем синонимии и полисемии слов. Вероятностная тематическая модель описывает каждую тему дискретным распределением на множестве терминов, каждый документ — дискретным распределением на множестве тем [2].

За последние 15 лет в литературе были описаны сотни разновидностей тематических моделей [6], включая модели жанровой классификации текстов, модели с частичным обучением, модели с автоматическим определением числа тем, модели с улучшенными показателями устойчивости и интерпретируемости тем, модели с учётом геопространственной мета-информации.

Современные модели, методы и алгоритмы тематического моделирования основаны на теории байесовского обучения и графических моделей. Вывод метода оптимизации для каждой модели является отдельной математической задачей, что влечёт большие затраты времени на разработку, реализацию и тестирование тематических моделей в прикладных проектах. Более того, байесовские методы не позволяют строить многофункциональные модели с заданными свойствами путём комбинирования известных моделей. Для решения данной проблемы в [14] предложен альтернативный подход к тематическому моделированию — аддитивная регуляризация тематических моделей (ARTM), основанный на классической не-байесовской теории регуляризации некорректно поставленных задач по А.Н.Тихонову. Показано, что он позволяет описывать широкий класс известных байесовских моделей с помощью регуляризаторов правдоподобия, а также комбинировать регуляризаторы, тем самым комбинируя модели [11]. Оптимизация любых моделей и их комбинаций производится одним и тем же регуляризированным EM-алгоритмом. При этом регуляризаторы не обязаны иметь вероятностную интерпретацию, как в байесовском подходе. Исследователь получает возможность сосредоточиться на формализации требований к модели, практически не заботясь о возможных математических и технических сложностях.

Простота и универсальность математического аппарата APTM позволила создать модульную технологию построения многокритериальных тематических моделей на основе библиотеки регуляризаторов. Каждый регуляризатор обеспечивает определённое свойство тематической модели. Для решения прикладной задачи пользователь выбирает из библиотеки набор регуляризаторов, обеспечивающий построение модели с требуемыми свойствами. Таким образом, технология APTM целиком исключает из процесса разработки модели этап решения нетривиальных математических проблем и открывает доступ к достижениям современного тематического моделирования для прикладных аналитиков.

2.1 Мультимодальные тематические модели

Обозначим через M множество модальностей. Каждая модальность $m \in M$ имеет словарь W^m, элементы которого называются токенами. Слова, составляющие основной текст документов, образуют первую модальность W^1. Если в документах заранее выделяются тэги, именованные сущности (named entity) или словосочетания (n-граммы), то они, как правило, рассматриваются как отдельные модальности. Объединение словарей всех модальностей обозначим через W. Каждый документ — это последовательность токенов различных модальностей. Предполагается, что тематика текста может быть выявлена по частотам токенов в документах, а порядок токенов не важен. Это предположение называют гипотезой «мешка слов». Итак, коллекция задаётся частотами (числом вхождений) n_{dw} токенов w в документах d. Вероятностная тематическая модель описывает вероятность появления токенов w модальности m в документах d распределением $p(w \mid d) = \sum_{t \in T} \phi^m_{wt} \theta_{td}$., где $\phi^m_{wt} = p(w \mid t)$ — представление темы t распределением вероятностей на множестве токенов W^m, $\Theta_{td} = p(t \mid d)$ — представление документа d распределением вероятностей на множестве тем T, общее для всех модальностей. Матрицы $\Phi^m = \|\phi^m_{wt}\|$ и $\Theta = \|\theta_{td}\|$ являются искомыми параметрами модели. Обозначим через Φ матрицу, составленную из матриц Φ^m, поставленных в столбец друг на друга. Для каждой модальности запишем задачу максимизации логарифма правдоподобия:

$$L_m(\Phi^m, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \;\to\; \max_{\Phi^m, \Theta} \tag{1}$$

$$\sum_{w \in W} \phi^m_{wt} = 1,\ \phi_{wt} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geq 0. \tag{2}$$

Число тем $|T|$ обычно много меньше числа документов $|D|$ и объёма словаря $|W|$. Таким образом, построение тематической модели сводится к разложению матрицы нормированных частот $p(w \mid d) = \frac{n_{dw}}{n_d}$ в произведение двух матриц меньшего размера Φ и Θ. Данная задача имеет в общем случае бесконечно много решений, то есть является некорректно поставленной (недоопределённой).

2.2 Аддитивно регуляризованные тематические модели

Согласно теории регуляризации А.Н.Тихонова, если задача недоопределена, то её решение можно сделать устойчивым, добавив к основному критерию дополнительный критерий-регуляризатор, учитывающий специфику задачи и знания предметной области. Аддитивная регуляризация тематических моделей — это многокритериальный подход, в котором к основному критерию логарифма правдоподобия добавляется взвешен-

ная сумма регуляризаторов
$R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$ [11]. Мультимодальная аддитивно регуляризованная тематическая модель строится путём максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов.

$$\sum_m \tau_m L_m(\Phi^m, \Theta) + \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta) \;\to\; \max_{\Phi,\Theta}.$$

при ограничениях неотрицательности и нормировки столбцов матриц Φ^m и Θ. Веса τ_m и τ_i называются коэффициентами регуляризации. Теорема [13]. Если функции R_i гладкие и (Φ, Θ) —решение данной оптимизационной задачи, то оно удовлетворяет следующей системе уравнений со вспомогательными переменными p_{tdw}, n_{wt}, n_{td}:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\phi_{wt}\theta_{td}\big); \tag{3}$$

$$\phi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right); \quad n_{wt} = \sum_{d\in D} \tau_{m(w)} n_{dw} p_{tdw}; \tag{4}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right); \quad n_{td} = \sum_{w\in d} \tau_{m(w)} n_{dw} p_{tdw}. \tag{5}$$

где $m(w)$ — модальность токена w, оператор $\underset{t\in T}{\operatorname{norm}}(x_t)$ преобразует произвольный вектор $\|x_t\|$ в дискретное распределение вероятностей путём обнуления отрицательных компонент вектора и последующей нормировки. Вспомогательные переменные $p_{tdw} = p(t \mid d, w)$ имеют ясную вероятностную интерпретацию — это условное распределение тем для каждого токена w в каждом документе d. Вспомогательные переменные n_{wt} — это счётчики числа употреблений токена w, связанных с темой t; n_{td} — счётчики числа употреблений токенов, связанных с темой t, в документе d. Для решения приведённой выше системы уравнений используется EM-алгоритм, который в данном случае совпадает с методом простых итераций. Каждая итерация — это один проход всей коллекции. Перед первой итерацией параметры ϕ_{wt}^m инициализируются случайными значениями, параметры θ_{td} — равномерными распределениями. На каждой итерации поочерёдно выполняются два шага: E-шаг (3) и M-шаг (5). На E-шаге вспомогательные переменные p_{tdw} вычисляются по основным параметрам модели ϕ_{wt} и θ_{td}. На M-шаге, наоборот, основные параметры ϕ_{wt} и θ_{td} вычисляются по вспомогательным переменным p_{tdw}. При рациональной организации вычислительного процесса значения p_{tdw} не хранятся, а вычисляются при необходимости. Вычислительная сложность EM-алгоритма линейна по объёму коллекции, по числу тем и по числу итераций. Он хорошо масштабируется, поскольку, число итераций, необходимых для сходимости процесса, как правило, невелико.

2.3 Примеры регуляризаторов

Множество тем T тематической модели разделим на две группы: подмножество этнорелевантных тем S и подмножество B всех остальных (фоновых) тем. К группам тем S и B регуляризаторы применяются по-разному.

2.3.1 Регуляризаторы разреживания и сглаживания

Сглаживание

Потребуем, чтобы распределения слов в темах ϕ_{wt} были близки к заданному распределению β_w, одинаковому для всех тем; а распределения тем в документах θ_{td} — к распределению α_t, одинаковому для всех документов. Будем минимизировать дивергенции Кульбака–Лейблера:

$$\sum_{t\in T}\mathrm{KL}(\beta\|\phi_t) = \sum_{t\in B}\sum_{w\in W}\beta_w \ln\frac{\beta_w}{\phi_{wt}} \;\to\; \min \tag{6}$$

$$\sum_{d\in TD}\mathrm{KL}(\alpha\|\theta_d) = \sum_{d\in D}\sum_{t\in B}\alpha_t \ln\frac{\alpha_t}{\theta_{td}} \;\to\; \min \tag{7}$$

Объединяя их в один критерий и отбрасывая слагаемые, не зависящие от параметров модели ϕ_{wt} и θ_{td}, получим регуляризатор сглаживания:

$$R(\Phi,\Theta) = \beta_0\sum_{t\in B}\sum_{w\in W}\beta_w \ln\phi_{wt} + \alpha_0\sum_{d\in D}\sum_{t\in B}\alpha_t\ln\theta_{td} \;\to\; \max$$

где β_0, α_0 — коэффициенты регуляризации. Тогда общая формула М-шага (5) даст те же оценки параметров, что и стандартная модель LDA [3]:

$$\phi_{wt} = \underset{w}{\mathrm{norm}}(n_{wt} + \beta_0\beta_w); \quad \theta_{td} = \underset{t}{\mathrm{norm}}(n_{dt} + \alpha_0\alpha_t)$$

Эффектом данного регуляризатора является сглаживание (увеличение) малых значений параметров ϕ_{wt} и θ_{td} за счёт незначительного уменьшения больших значений. Регуляризатор сглаживания применяется к фоновым темам B, что способствует переносу слов общей лексики из этнорелевантных тем в фоновые.

Разрежевание

Естественно предполагать, что каждый документ относится к малому числу тем, и каждая тема характеризуется относительно небольшим набором терминов. В таком случае большинство параметров ϕ_{wt} и θ_{td} должны быть равны нулю. В ARTM требование разреженности формализуется просто — с помощью регуляризатора, максимизирующего дивергенцию Кульбака–Лейблера между распределениями ϕ_{wt} и β_w, а также θ_{td} и α_t,

$$R(\Phi,\Theta) = -\beta_0\sum_{t\in S}\sum_{w\in W}\beta_w \ln\phi_{wt} - \alpha_0\sum_{d\in D}\sum_{t\in S}\alpha_t\ln\theta_{td} \;\to\; \max$$

что приводит к формуле М-шага:

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} - \beta_0\beta_w); \quad \theta_{td} = \underset{t}{\text{norm}}(n_{dt} - \alpha_0\alpha_t)$$

Отметим, что сама возможность описания сглаживания и разреживания общей формулой с единственным различием в знаке параметра до сих оставалась незамеченной в байесовском подходе, поскольку требовала отказаться от интерпретации регуляризатора через априорное распределение Дирихле.

2.3.2 Регуляризатор декоррелирования тем

Известно, что повышение различности тем улучшает интерпретируемость модели [5]. Чтобы темы как можно сильнее различались, вводится регуляризатор, определяемый через сумму ковариаций между распределениями ϕ_{wt} и ϕ_{ws} для всех пар тем t, s [10]:

$$R(\Phi, \Theta) = -\tau \sum_{t,s \in T} \sum_{w \in W} \phi_{wt}\phi_{ws} \ \rightarrow \ \max$$

где τ — коэффициент регуляризации. Это приводит к формуле М-шага

$$\phi_{wt} = \underset{w}{\text{norm}}(n_{wt} - \tau\phi_{wt} \sum_{s \in T, s \neq t} \phi_{ws})$$

Декоррелирование применяется к этнорелевантным темам и приводит к их разреживанию и более чёткому выделению лексических ядер, состоящих из характерных слов w с сильно доминирующей вероятностью $\phi_{wt} = p(w \mid t)$ в данной теме t. Декоррелирование, как и разреживание, хорошо сочетается со сглаживанием фоновых тем. В этом случае эффектом декоррелирования становится выделение относительно небольшого ядра в каждой теме, содержащего, согласно экспериментам [10], от 30 до 200 слов.

3 Задача классификации

В работе решается задача классификации на два класса. В качестве объектов выступает множество тем построенных тематических моделей. Приведем формальную постановку задачи.

Пусть X — множество признаковых описаний объектов, $Y = 0, 1$ — множество номеров (имён, меток) классов. Существует неизвестная целевая зависимость — отображение $y^\star : X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), ..., (x_m, y_m)\}$. Требуется построить алгоритм $a : X \to Y$, способный классифицировать произвольный объект $x \in X$.

3.1 Оценивание качества бинарной классификации

Многие алгоритмы классификации, в частности, использованные в данной работе, могут возвращать вероятность принадлежности классифицируемого объекта к положительному классу. Варьируя порог отсечения вероятности, можно получать то или иное разбиение классифицируемых объектов на два класса.

- TP (True Positives) – верно классифицированные положительные примеры (так называемые истинно положительные случаи)
- TN (True Negatives) – верно классифицированные отрицательные примеры (истинно отрицательные случаи)
- FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый "ложный пропуск"– когда интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры)
- FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи)

При анализе чаще оперируют не абсолютными показателями, а относительными:

- Доля истинно положительных объектов, полнота (True Positives Rate):

$$TPR = \frac{TP}{TP + FN}$$

- Доля ложно положительных объектов (False Positives Rate):

$$FPR = \frac{FP}{TN + FP}$$

- Доля истинно положительных объектов среди предсказанных положительными:

$$Precision = \frac{TP}{TP + FP}$$

ROC-кривая — графическая характеристика качества бинарного классификатора, зависимость доли стинно положительных классификаций (TPR) от доли ложно положительных классификаций (FPR) при варьировании порога решающего правила. Площадь под данной кривой (ROC AUC) показывает прогностическую силу модели.

Cumulative Gain – CG_p показывает долю релевантных (положительных) объектов в первых p объектах выборки, отсортированной в порядке убывания вероятности отнесения к положительному классу.

$$CG_p = \frac{\sum_{i=0}^{p} rel_i}{\sum rel_i}$$

где rel_i – индикатор релевантности класса.

Lift – показывает во сколько раз доля положительных объектов в первых p объектах выборки, отсортированной в порядке убывания вероятности отнесения к положительному классу, больше доли положительных объектов в p объектах, взятых случайным образом. Равен отношению CG_p к доле релевантных объектов среди p случайно выбранных.

4 Преобразование признакового пространства

4.1 Структура данных

В рассматриваемой задаче классификации, в качестве объектов обучения выступают темы некоторой построенной тематической модели. Объект-тема x_t является столбцом матрицы Ф:

$$x_t = \phi_t^T = [p(w_1|t), \ldots, p(w_{|w|}|t)]$$

и представляет из себя дискретное распределение, т.е. числовой вектор длины $|W|$.
Чаще всего размер словаря составляет несколько десятков тысяч слов, а количество размеченных тем ограничено возможностями асессора и исчисляется сотнями. Поэтому особенность данной задачи состоит в многократном преобладании количества признаков над количеством объектов обучающей выборки.

4.2 Мета-признаки

При анализе задачи возникает естественное предположение о том, что если тема близка к другим темам класса C и далека от тем остальных классов, то, скорее всего, она сама принадлежит классу C. На основе этого предположения были опробованы различные подходы, так или иначе опирающиеся на различные меры близости между темами.

Для каждого объекта обучающей выборки строим расстояния до ближайшего объекта каждого класса.

Используемые виды расстояний (M – число признаков):

- KL-дивергенция (симметризованная) $d_{KL}(p,q) = \mathrm{KL}(p\|q) + \mathrm{KL}(q\|p)$
- Евклидово расстояние $d_e(p,q) = \sqrt{\sum_{i=1}^{M}(q_i - p_i)^2}$
- Косинусное расстояние $d_{cos}(p,q) = 1 - \frac{(p,q)}{\|p\|\|q\|}$
- Расстояние Хеллингера $d_{hel}(p,q) = \frac{1}{\sqrt{(2)}} d_e(\sqrt{p}, \sqrt{q})$

.

Расстояния вычислялись при помощи процедуры Leave-one-out кросс-валидации. Т.е. для каждого объекта считаем p различных расстояний до ближайших объектов каждого из l классов, усредняя по $1 \dots k$ соседям. Итого, получаем $m = pkl$ признаков.

4.3 Преобразование исходных признаков

Конечно, можно решать задачу классификации на оригинальных данных без построения мета-признаков. Однако, здесь мы сталкиваемся с проблемой высокой размерности данных, по отношению к количеству обучающих объектов. Частично эта проблема решается усечением распределения.

Число реально относящихся к теме терминов обычно много меньше размера всего словаря. Для того, чтобы убрать из темы лишние слова, используется обнуление слов с малой вероятностью, называемое усечением $p(w|t)$:

$$p(w|t)\,[p(w|t) < \alpha] = 0$$

В качестве порога уместно выбрать значение $\alpha = \frac{1}{|W|}$.

После применения процедуры усечения, можно ввести новое расстояние для построения мета-признаков, также называемое расстоянием Хэмминга:

$$p_{ham}(p,q) = \frac{1}{M}\sum_{i=1}^{M}[[p_i > 0] = [q_i > 0]]$$

Также можно применить различные преобразования исходных признаков, например извлечение квадратного корня:

$$p(w|t) = \sqrt{p(w|t)}$$

5 Эксперименты

5.1 Данные

Для обучения задачи классификации использовались тематические модели, построенные авторами статьи [1]. Данные тематические модели построены по двум коллекциям: LiveJournal и IQBuzz.

LiveJournal. Коллекция представляет собой примерно 1.58 млн. лемматизированных постов сервиса. Словарь коллекции составил 860 тыс. слов. После предобработки, заключавшейся в удалении всех слов, содержащих что-либо, кроме символов кириллицы (с одним опциональным дефисом), имеющих длину менее 3 символов и встречающихся реже 20 раз во всей коллекции, словарь сократился до 90 тысяч. Удаление всех пустых документов привело к тому, что итоговая длина коллекции составила около 1.38 млн. документов.
Все модели были обучены онлайновым алгоритмом библиотеки BigARTM – библиотеке тематического моделирования с открытым кодом [12]. В ней реализован наиболее эффективный на сегодняшний день онлайновый параллельный EM-алгоритм для тематического моделирования больших коллекций, а также имеется встроенная расширяемая библиотека регуляризаторов и метрик качества тематических моделей.
В каждой модели использовалось оптимальное число тем $|T| = 400$, причем в моделях с регуляризацией множество тем делилось на $|B| = 150$ фоновых и $|S| = 250$ предметных.

Рассмотрим использованные модели (в скобках указано кодовое название модели, используемое далее в таблицах):

- Модель вероятностного латентного семантического анализа PLSA, в которой регуляризаторы отсутствуют. (**plsa**)
- Модель латентного размещения Дирихле LDA, реализованная в BigARTM как модель с регуляризаторами сглаживания столбцов матриц Φ и Θ по равномерным распределениям. (**lda**)
- Модель APTM с регуляризаторами сглаживания и разреживания по словарю этнонимов. Кроме этого, в этой и всех последующих регуляризованных моделях применялись регуляризаторы сглаживания матрицы Θ по фоновым темам, и разреживания по предметным темам. (**smooth**)
- Модель APTM, улучшение предыдущей модели путём добавления регуляризатора декоррелирования. (**smooth_decor**)

- Модель со всеми регуляризаторами, использованными выше, в которую добавлена отдельная модальность этнонимов. (**full**)

- Модель PLSA без регуляризаторов, обученная на документах, ограниченных темами модели full, которые эксперты признали этнорелевантными. В обучение вошли документы, значение которых в матрице Θ выше некоторого порога. (**plsa_ethnic**)

IQBuzz. Коллекция содержит около 6 млн лемматизированных сообщений различных социальных медиа. Основные из них: ВКонтакте, Twitter, Google+, LiveInternet, ursatm.ru, Эхо Москвы. Словарь сырой коллекции составил 8.3 млн слов, но после глубокой фильтрации и предобработки сократился до 75 тысяч.

Тематические модели для данной коллекции также были получены с помощью онлайнового алгоритма BigARTM [12]:

- Модель PLSA без регуляризаторов. (**iq_plsa**)

- Модель APTM с регуляризатором модальностей этнонимов и биграмм с этнонимами [1]. (**iq_artm**)

Была произведена экспертная разметка полученных тем на 2 класса:

- имеющие отношение к межэтническим и международным отношениям (y = 1)

- нерелевантные темы (y = 0)

Итого, получены следующие выборки для задачи классификации:

- **LiveJournal**, X_{lj}
  2400 обучающих объектов, среди которых 2141 относятся к классу $y = 0$ и 259 объектов класса $y = 1$.

- **IQBuzz**, X_{iq}
  400 обучающих объектов, среди которых 314 относятся к классу $y = 0$ и 86 объектов класса $y = 1$.

5.2 Валидация

Для оценивания качества классификации проводилась процедура скользящего контроля по моделям, устроенная следующим образом.

Каждый объект-тема x_t принадлежит одной из тематических моделей:

$$x_t \in M_i, i = 1 \dots 8$$

(6 моделей построенных по коллекции LiveJournal и 2 модели IQBuzz).

Обучающая выборка X делилась непересекающиеся блоки M_i так, что в каждый блок входят объекты-темы, соответствующие одной модели:

$$X = \sqcup_i M_i, \quad \sum_k |M_k| = |X|$$

Для оценки качества модели выборка последовательно разбивается на 2 части: обучающую $X \backslash M_k$ и тестовую M_k. На каждом шаге происходит обучение классификатора на обучающей выборке и предсказание ответов для тестовой. Таким образом получаются оценки качества классификации каждой тематической модели M_i. Для того, чтобы получить общую оценку качества среди всех моделей, предсказания, полученные для каждого разбиения, конкатенируются в единый вектор предсказаний a, после чего вычисляются метрики качества.

Также интерес представляет следующий вопрос: может ли классификатор, обученный на тематических моделях, построенных по одной коллекции хорошо предсказывать классы тем из тематических моделей другой коллекции? И насколько будет отличаться качество предсказания на новой коллекции?

Для ответа на эти вопросы предлагается использовать три варианта проведения валидации:

- Скользящий контроль по выборке $X_l j$ (LJ_CV)
- Скользящий контроль по объединенной выборке $X_{lj} \bigsqcup X_{iq}$ (LJ_IQ_CV)
- Обучение на X_{lj}, валидация на X_{iq} (IQ_val)

В первых двух вариантах используется процедура скользящего контроля, описанная выше. Обучение только на выборке IQBuzz не проводилось, из-за малого количества объектов.

Также, если в обучении или предсказании участвует новая модель, необходимо использовать общий словарь терминов. При пересечении словаря двух коллекций, его объем уменьшился до $|W| = 40$ тыс. слов.

5.3 Алгоритмы

Для обучения использовались различные комбинации преобразований признаков и самих обучающих алгоритмов. Далее приведен их список с условным обозначением в скобках, которые используются в таблицах и графиках.

Виды признаков

1. Оригинальные $p(w|t)$ (raw)
2. Усеченные признаки (cut)
3. Преобразование $\sqrt{p(w|t)}$ (sqrt)
4. Преобразование $\sqrt{p(w|t)}$ после усечения (cut_sqrt)
5. Мета-признаки (meta)

Преобразование $\sqrt{p(w|t)}$ было выбрано экспериментальным путем, также оно имеет косвенное отношение к метрике Хеллингера $d_{hel}(p,q) = \frac{1}{\sqrt{(2)}} d_e(\sqrt{p}, \sqrt{q})$, которая неплохо себя показала в качестве мета-признаков.

Алгоритмы

1. Логрегрессия (log)
2. Логрегрессия с L2 регуляризацией (rdg)
3. Линейный SVM (svc)
4. Градиентный бустинг над решающими деревьями (xgb)
5. Случайный лес (rf)

Данный список включает в себя не все алгоритмы машинного обучения, с которыми проводились эксперименты, однако лучших результатов удалось достичь именно с ними.

Подробное теоретическое описание каждого метода можно найти в [8]. В работе использовались следующие реализации алгоритмов:

- LIBLINEAR [7] для логистической регрессии и SVM
- Scikit-learn [9] для случайного леса
- XGBoost [4] для градиентного бустинга

5.4 Результаты

В работе были обучены модели, использующие различные пары преобразований признаков и алгоритмов машинного обучения. Гиперпараметры алгоритмов настраивались по скользящему контролю на выборке LiveJournal.

Общие результаты метрики ROC AUC в разделении по способу валидации представлены в Таблице 1. Лучшим алгоритмом по качеству валидации внутри выборки LiveJournal является линейный SVM с преобразованием признаков cut_sqrt. Однако на валидации внутри объединенной выборки и при тестировании на отдельной выборке IQBuzz лучшее качество показала логистическая регрессия с L2 регуляризацией на тех же признаках. Она же показывает лучшее усредненное по способам валидации качество. Графики Lift, CG и ROC лучшей модели показаны на Рис. 1.

Model	LJ_CV	LJ_IQ_CV	IQ_val
log	0.8409	0.8353	0.8260
svc	0.8210	0.8161	0.7894
rdg	0.8026	0.7896	0.6622
sqrt_log	0.9482	0.9210	0.8355
sqrt_svc	0.9470	0.9194	0.8311
sqrt_rdg	0.9460	0.9211	0.8366
cut_log	0.9225	0.8955	0.8279
cut_svc	0.9153	0.8739	0.8055
cut_rdg	0.9283	0.9043	0.8250
cut_sqrt_log	0.9501	0.9232	0.8384
cut_sqrt_svc	**0.9508**	0.9243	0.8436
cut_sqrt_rdg	**0.9471**	**0.9250**	**0.8467**
log_meta	0.9149	0.9218	0.8282
svc_meta	0.9156	0.9220	0.8277
rdg_meta	0.9151	0.9212	0.8285
rf_meta	0.9017	0.9022	0.8359
xgb_meta	0.9153	0.9075	0.8372

Таблица 1: ROC AUC по различным методам валидации.

Например, из графика Lift на Рис. 1 следует, что при валидации на выборке LiveJournal в топ-20% объектов, отсортированных по убыванию предсказанной вероятности класса $y = 1$ находится примерно в 5 раз больше релевантных объектов, чем в 20% выбор-

ки, выбранных случайно. А для предсказаний на новой выборке IQBuzz это значение значительно меньше — в топ 20% находится в 3 раза больше релевантных объектов.

Основной вывод — качество предсказания при обучении и тесте на тематических моделях, построенных по разным коллекциям значительно меньше, чем качество полученное внутри одной коллекции.

Заметим, что качество классификации больше зависит от типа преобразования признаков, чем от конкретного алгоритма обучения. На Рис. 2, 3, 4 представлены графики Lift, CG и ROC лучшего алгоритма обучения в разбиении по группам преобразований признаков.

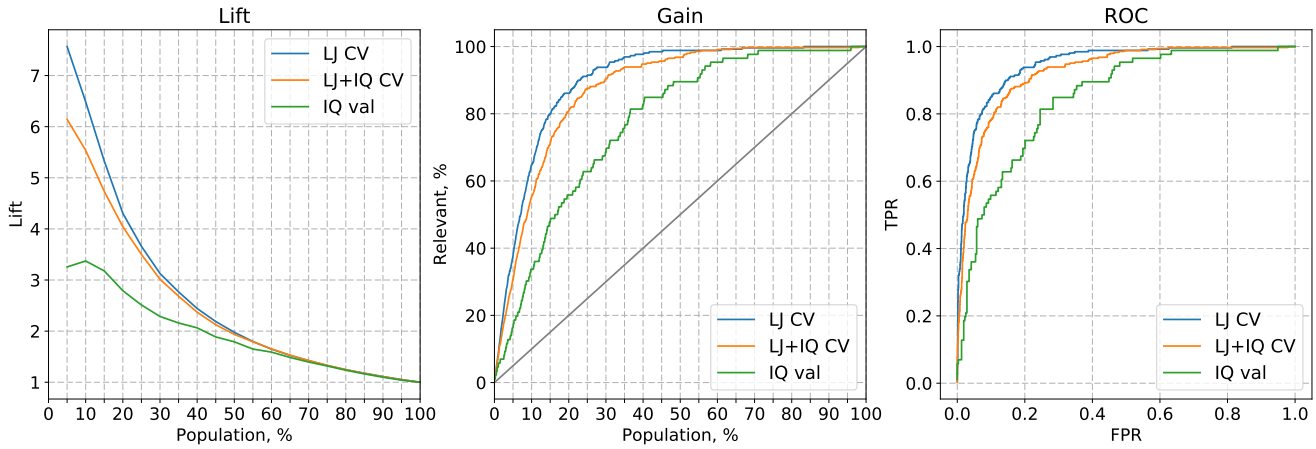


Рис. 1: Логистическая регрессия с L2-регуляризацией на усеченных признаках с преобразованием $\sqrt{p(w|t)}$

Рассмотрим результаты классификации отдельно по способом валидации.

В Таблице 2 представлены результаты скользящего контроля на выборке LiveJournal при тестировании по отдельным тематическим моделям. Аналогично, в Таблице 3 представлены результаты валидации в объединенной выборке, а и в Таблице 4 результаты валидации на новой выборке IQBuzz.

Можно сделать вывод, что структура тем некоторых тематических моделей сильно отличается от остальных. Например, хуже всего предсказываются темы модели smooth_decor. Также, алгоритмы, построенные на мета-признаках, хорошо предсказывают классы тем модели smooth_decor, обгоняя по качеству многие алгоритмы на признаках cut_sqrt. А для тематической модели full, наоборот, мета-признаки проигрывают всем остальным.

Дополнительно, графики метрик в разбиении по алгоритмам и способу преобразования признаков находятся в Приложении.

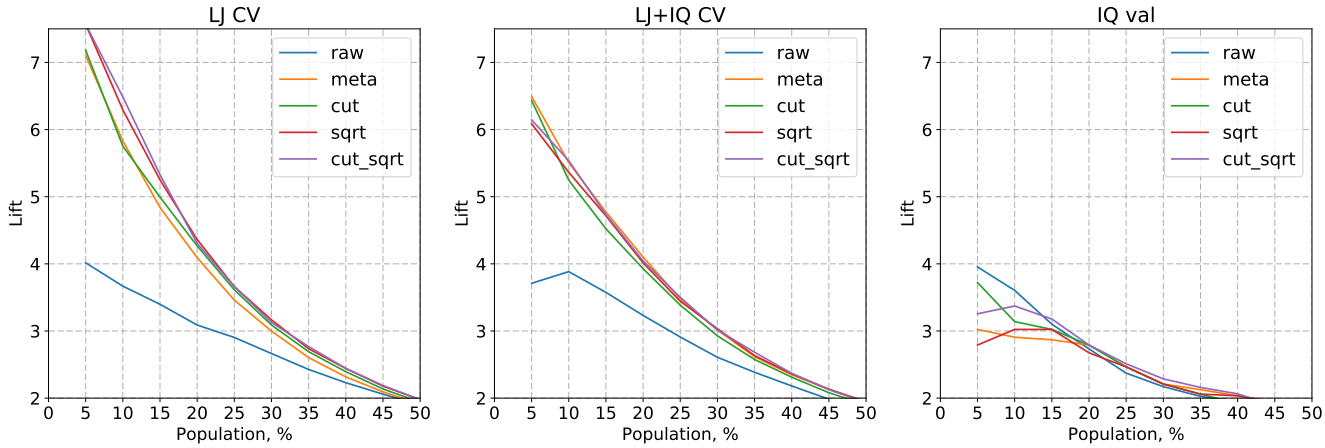


Рис. 2: Lift лучших алгоритмов обучения в группе преобразований признаков.

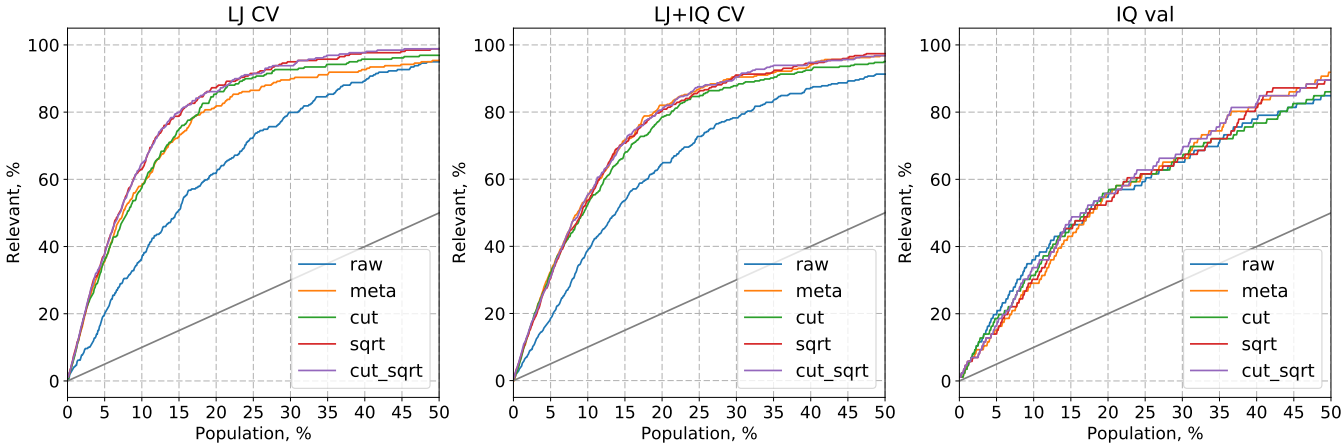


Рис. 3: CG лучших алгоритмов обучения в группе преобразований признаков.

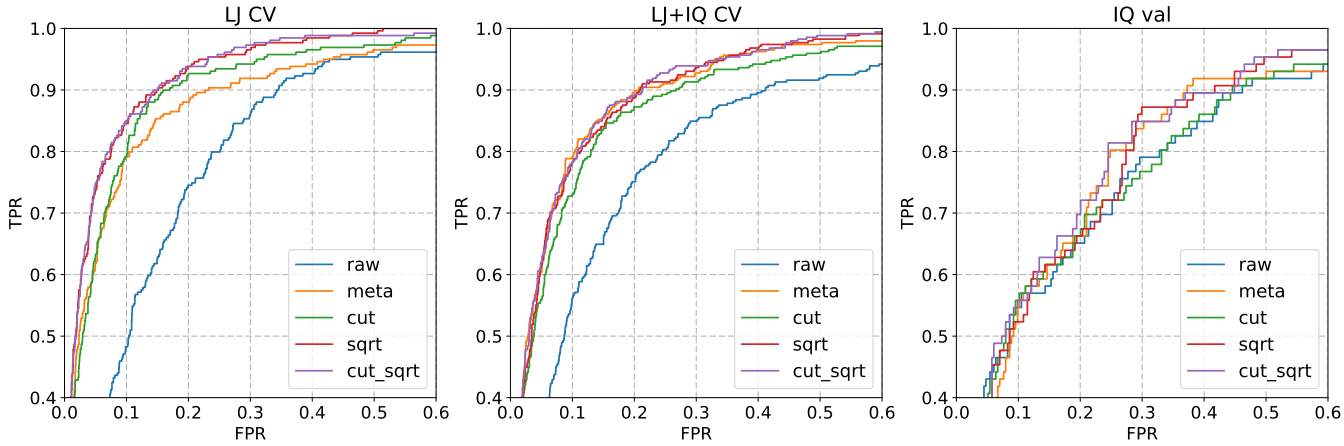


Рис. 4: ROC лучших алгоритмов обучения в группе преобразований признаков.

При обучении и тестировании на разных выборках, вместе с уменьшением значений метрик качества также сильно уменьшается их разброс между видами преобразований признаков. Также при такой модели валидации можно отметить более стабильное поведение качества при использовании мета-признаков.

Model	full	lda	plsa	plsa ethnic	smooth	smooth decor	overall
log	0.8133	0.8803	0.9085	0.8864	0.8345	0.8379	0.8409
svc	0.7968	0.8800	0.8786	0.8747	0.8203	0.8202	0.8210
rdg	0.7722	0.8923	0.8537	0.8145	0.7948	0.7859	0.8026
sqrt_log	0.9486	0.9720	0.9687	0.9637	0.9473	0.8823	0.9482
sqrt_svc	0.9457	0.9724	0.9659	**0.9694**	**0.9557**	0.8712	0.9470
sqrt_rdg	0.9488	0.9736	0.9748	0.9605	0.9343	0.8705	0.9460
cut_log	0.9041	0.9645	0.9719	0.9252	0.9239	**0.8965**	0.9225
cut_svc	0.9144	0.9650	0.9750	0.9292	0.9411	0.8613	0.9153
cut_rdg	0.9143	0.9668	**0.9758**	0.9296	0.9224	0.8956	0.9283
cut_sqrt_log	0.9476	**0.9756**	0.9679	0.9660	0.9523	0.8847	0.9501
cut_sqrt_svc	**0.9493**	**0.9753**	**0.9713**	**0.9683**	**0.9465**	**0.8802**	**0.9508**
cut_sqrt_rdg	0.9469	0.9749	0.9749	0.9604	0.9335	0.8735	0.9471
log_meta	0.8797	0.9566	0.9724	0.9262	0.8720	0.8893	0.9149
svc_meta	0.8807	0.9566	0.9724	0.9257	0.8746	0.8889	0.9156
rdg_meta	0.8887	0.9581	0.9712	0.9225	0.8873	0.8870	0.9151
rf_meta	0.8683	0.9371	0.9498	0.9315	0.8758	0.8561	0.9017
xgb_meta	0.8765	0.9398	0.9644	0.9454	0.9244	0.8683	0.9153

Таблица 2: ROC AUC. Скользящий контроль по тематическим моделям. Валидация по выборке Livejournal.

Model	artm iq	full	lda	plsa	plsa ethnic	plsa iq	smooth	smooth decor	overall
log	0.8359	0.8338	0.8651	0.8814	0.8663	0.8562	0.8703	0.8167	0.8353
svc	0.8165	0.8123	0.8402	0.8502	0.8447	0.8482	0.8558	0.8009	0.8161
rdg	0.7700	0.7863	0.8482	0.8366	0.8299	0.8402	0.8337	0.7310	0.7896
sqrt_log	0.8516	**0.9443**	0.9763	0.9657	0.9596	0.9129	0.9582	0.8698	0.9210
sqrt_svc	0.8419	0.9362	0.9748	0.9613	0.9665	0.8906	0.9625	0.8588	0.9194
sqrt_rdg	0.8597	0.9374	0.9764	0.9703	0.9544	0.9193	0.9482	0.8570	0.9211
cut_log	0.7976	0.8965	0.9605	0.9664	0.9281	0.9264	0.9475	0.8816	0.8955
cut_svc	0.7547	0.9082	0.9725	0.9710	0.9332	0.9143	0.9593	0.8561	0.8739
cut_rdg	0.7985	0.9024	0.9691	0.9688	0.9297	0.9271	0.9517	0.8844	0.9043
cut_sqrt_log	0.8489	0.9424	0.9777	0.9633	0.9631	0.9105	0.9613	0.8740	0.9232
cut_sqrt_svc	0.8512	0.9387	0.9785	0.9676	0.9635	0.9103	0.9560	0.8689	0.9243
cut_sqrt_rdg	**0.8616**	**0.9349**	**0.9775**	**0.9699**	**0.9537**	**0.9190**	**0.9470**	**0.8661**	**0.9250**
log_meta	0.8107	0.8866	0.9725	0.9761	0.9296	0.9309	0.9240	0.9020	0.9218
svc_meta	0.8097	0.8862	0.9724	0.9762	0.9291	0.9318	0.9241	0.9020	0.9220
rdg_meta	0.8122	0.8891	0.9723	0.9753	0.9265	0.9363	0.9289	0.8992	0.9212
rf_meta	0.8183	0.8633	0.9468	0.9761	0.9124	0.9259	0.9078	0.8618	0.9022
xgb_meta	0.8371	0.8787	0.9494	0.9726	0.9164	0.9176	0.9338	0.8747	0.9075

Таблица 3: ROC AUC. Скользящий контроль по тематическим моделям. Валидация по объединенной выборке Livejournal и IQBuzz.

Model	artm_iq	plsa_iq	overall
log	0.8177	0.8501	0.8260
svc	0.7740	0.8248	0.7894
rdg	0.6644	0.6577	0.6622
sqrt_log	0.8269	0.8883	0.8355
sqrt_svc	0.8247	0.8783	0.8311
sqrt_rdg	0.8226	0.8880	0.8366
cut_log	0.7776	0.9036	0.8279
cut_svc	0.7404	0.8994	0.8055
cut_rdg	0.7716	0.9051	0.8250
cut_sqrt_log	0.8301	0.8864	0.8384
cut_sqrt_svc	0.8296	0.8930	0.8436
cut_sqrt_rdg	**0.8297**	**0.8935**	**0.8467**
log_meta	0.7729	0.8764	0.8282
svc_meta	0.7719	0.8771	0.8277
rdg_meta	0.7730	0.8769	0.8285
rf_meta	0.7977	0.8626	0.8359
xgb_meta	0.8017	0.8667	0.8372

Таблица 4: ROC AUC. Скользящий контроль по тематическим моделям. Валидация на выборке IQBuzz, обучение на Livejournal

Из приведенных выше таблиц и графиков можно сделать вывод, что использование усечения заметно улучшает качество классификации. Ранее был введен жесткий порог $\alpha = \frac{1}{|W|}$, значения признакого меньше которого обнуляются. Для проверки возможностей улучшения качества, был проведен эксперимент по оценке качества классификации лучшего алгоритма на валидации по выборке LiveJournal в зивисимости от порога усечения. Результаты приведены на Рис. 6. Из графика следует, что исходный выбор порога обоснован и практически совпадает с порогом, обеспечивающим максимум классификации.

Распределение количества ненулевых вероятностей слов $p(w|t)$ после усечения в объектетеме представлены на Рис. 5.

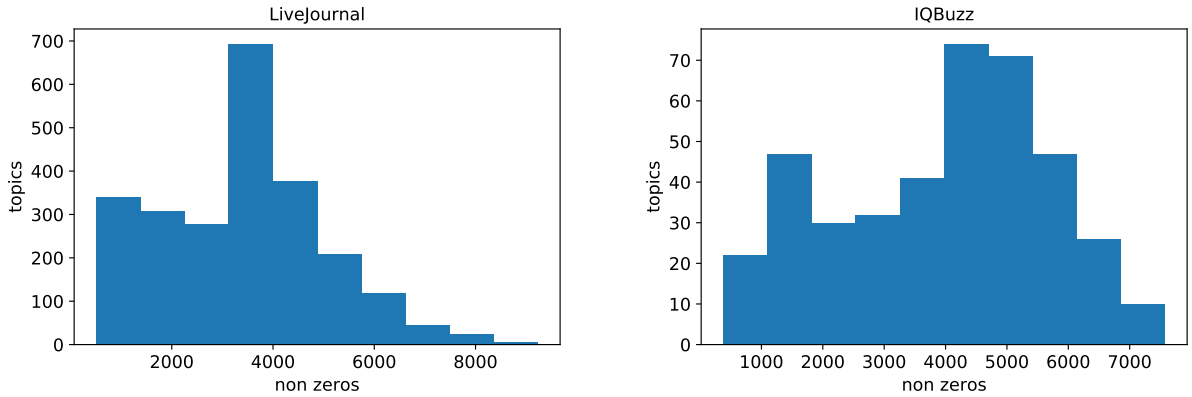


Рис. 5: Гистограммы распределения количества нулевых признаков.

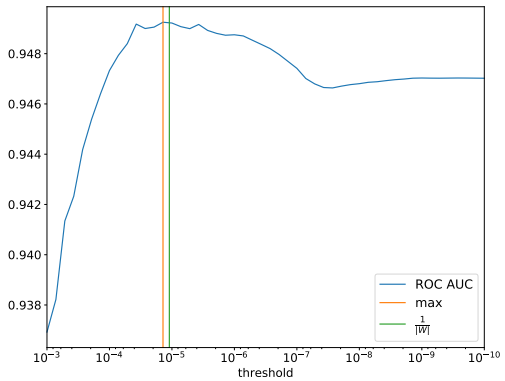


Рис. 6: Зависимость ROC AUC от порога усечения.

6 Заключение

В данной работе решалась задача автоматической классификации тем, построенных тематической моделью.

Было рассмотрено несколько подходов к формированию признакового пространства, в частности, было предложено уменьшить размерность данных, построив с помощью метода ближайших соседей мета-признаки, основанные на различных видах расстояний. Однако, эксперименты на реальных данных показали, что задача более точно решается в признаковом пространстве вероятностей терминов, несмотря на то, что число признаков многократно превышает объём обучающей выборки.

Было показано, что задача решается с довольно высокой точностью, тем самым значительно снижает трудозатраты на ручной анализ тем.

Также установлено, что классификатор, обученный на темах, построенных на одной текстовой коллекции, может достаточно хорошо предсказывать классы тем, построенных на другой текстовой коллекции.

Список литературы

[1] Koltsova O. Nikolenko S. Vorontsov K. Apishev M., Koltcov S. Mining ethnic content online with additively regularized topic models. *Computacion y Sistemas*, 20:387—-403, 2016.

[2] David M. Blei. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April 2012.

[3] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[4] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[5] Weiwei Cui, Shixia Liu, Li Tan, Conglei Shi, Yangqiu Song, Zekai Gao, Huamin Qu, and Xin Tong. Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2412–2421, December 2011.

[6] Ali Daud, Juanzi Li, Lizhu Zhou, and Faqir Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. *Frontiers of Computer Science in China*, 4(2):280–301, 2010.

[7] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[8] Trevor J. Hastie, Robert John Tibshirani, and Jerome H. Friedman. *The elements of statistical learning : data mining, inference, and prediction*. Springer series in statistics. Springer, New York, 2009. Autres impressions : 2011 (corr.), 2013 (7e corr.).

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[10] K. V. Vorontsov and A. A. Potapenko. Tutorial on probabilistic topic modeling: Additive regularization for stochastic matrix factorization. *Communications in Computer and Information Science*, 436:29–46, 2014.

[11] K. V. Vorontsov and A. A. Potapenko. Additive regularization of topic models. *Machine Learning*, 101(1):303–323, 2015.

[12] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. *BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections*, pages 370–381. Springer International Publishing, Cham, 2015.

[13] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, Marina Suvorova, and Anastasia Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. In *Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications*, TM '15, pages 29–37, New York, NY, USA, 2015. ACM.

[14] К. В. Воронцов. Аддитивная регуляризация тематических моделей коллекций текстовых документов. *Доклады Академии наук*, 455(3):268–271, 2014.

7 Приложение

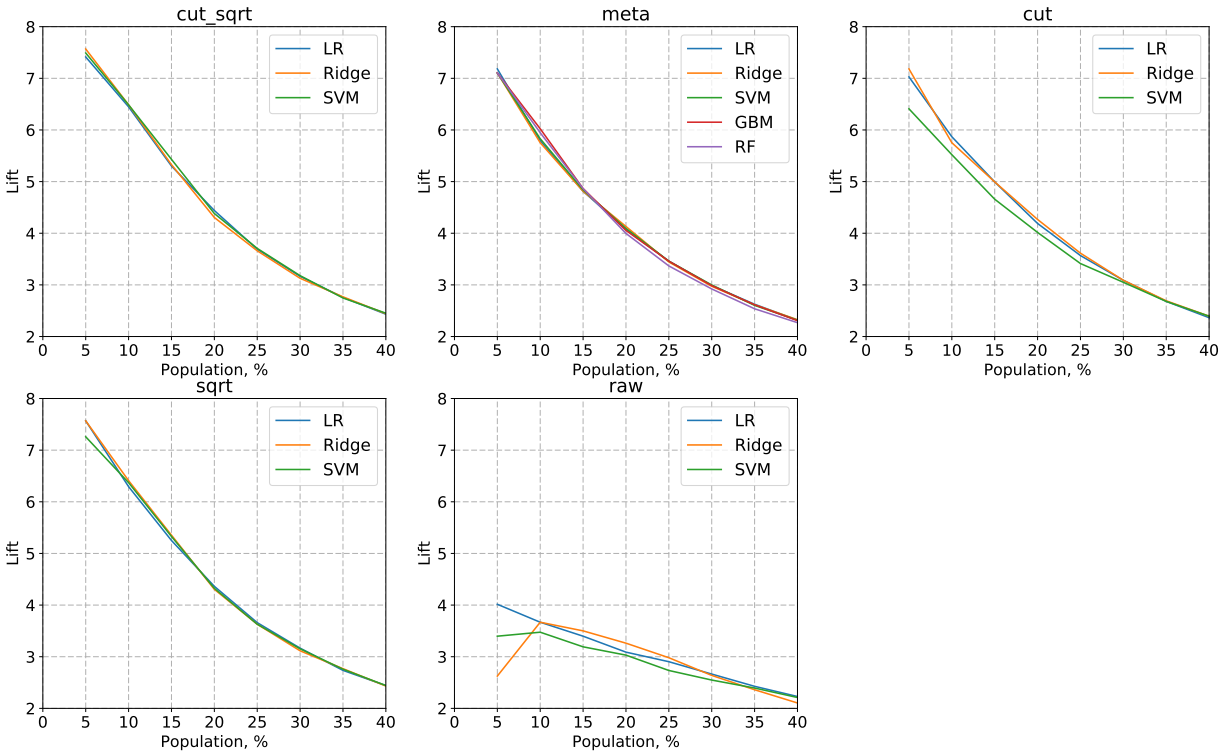


Рис. 7: Lift в разбиении по алгоритмам и преобразованию признаков.
Валидация по выборке Livejournal.

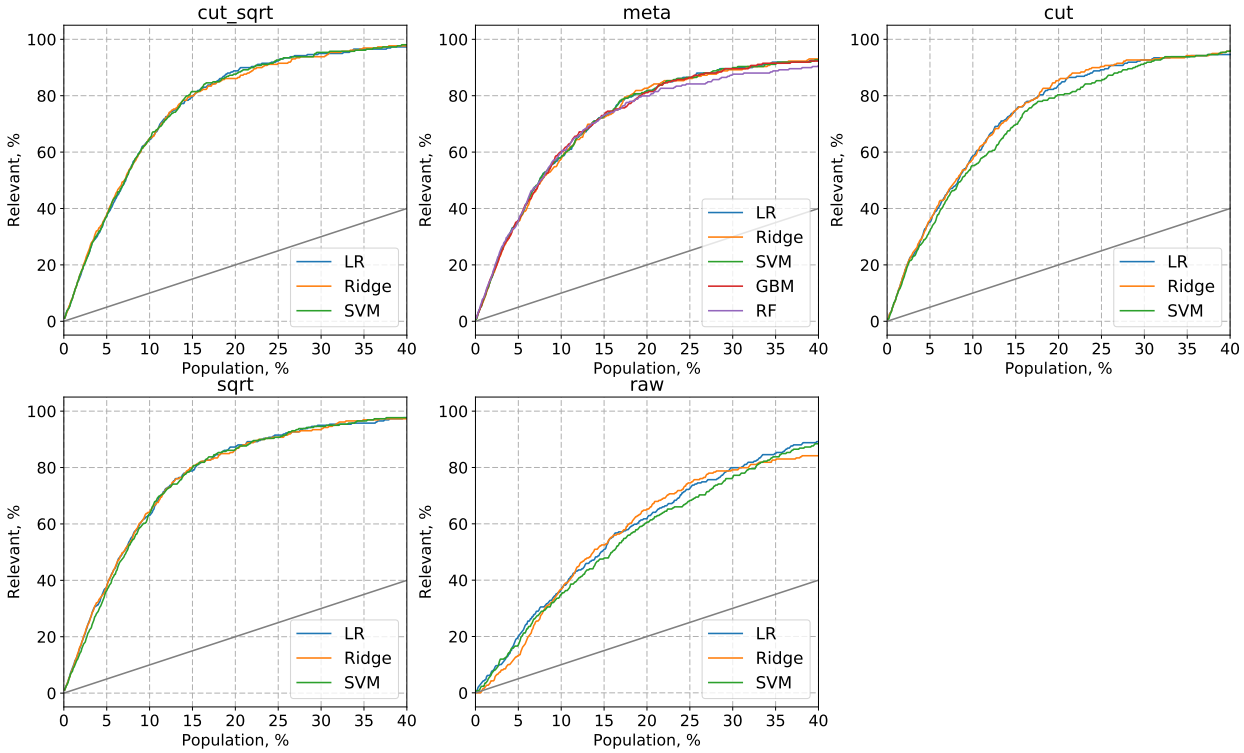


Рис. 8: CG в разбиении по алгоритмам и преобразованию признаков.
Валидация по выборке Livejournal.

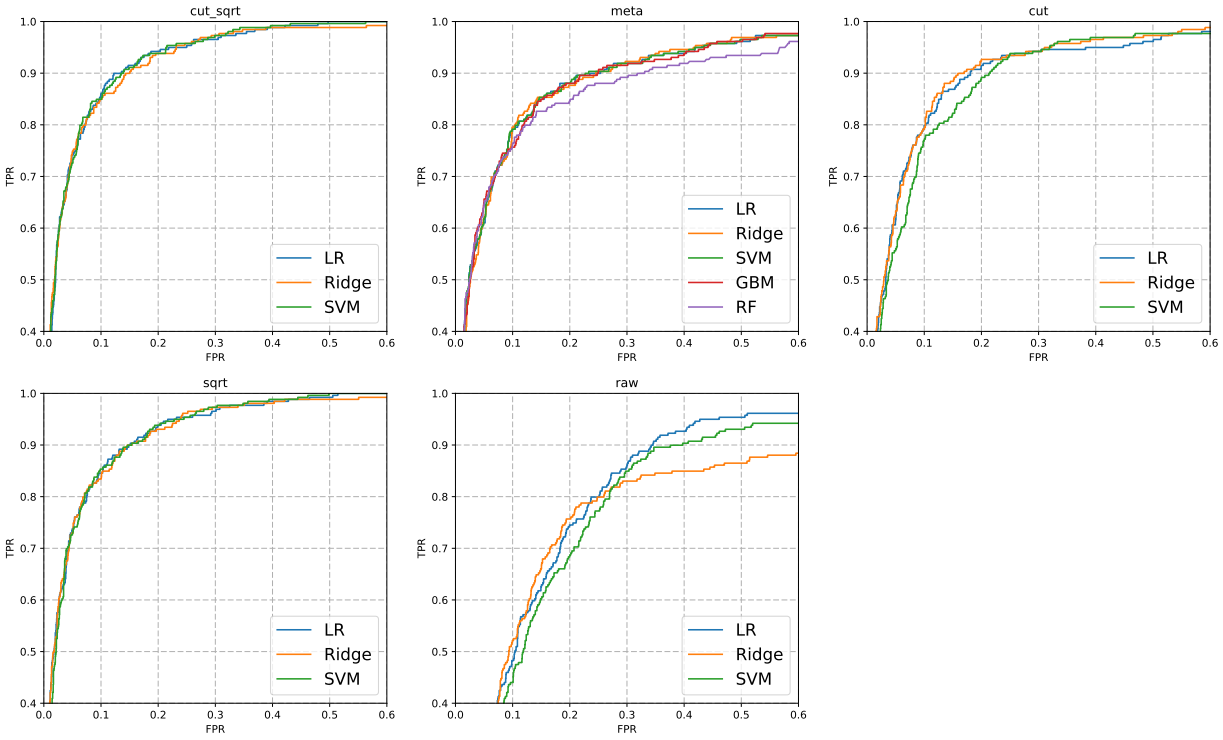


Рис. 9: ROC в разбиении по алгоритмам и преобразованию признаков. Валидация по выборке Livejournal.

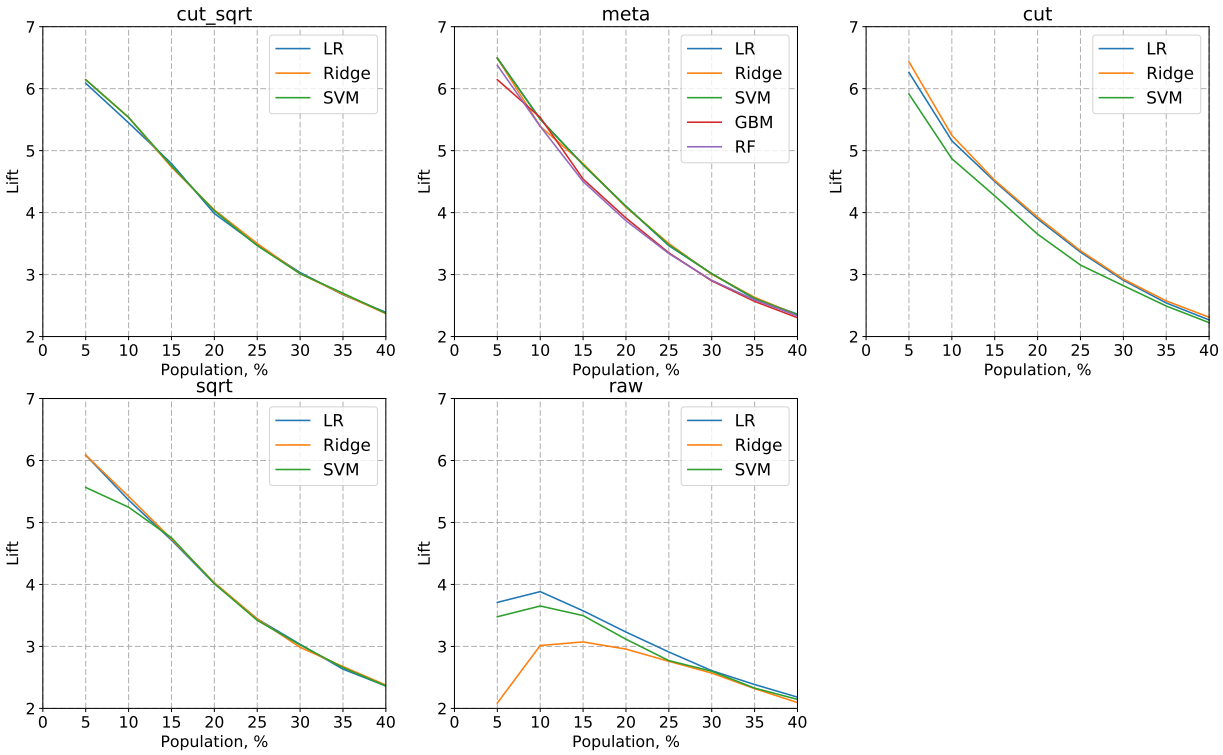


Рис. 10: Lift в разбиении по алгоритмам и преобразованию признаков.
Валидация по объединенной выборке Livejournal и IQBuzz.

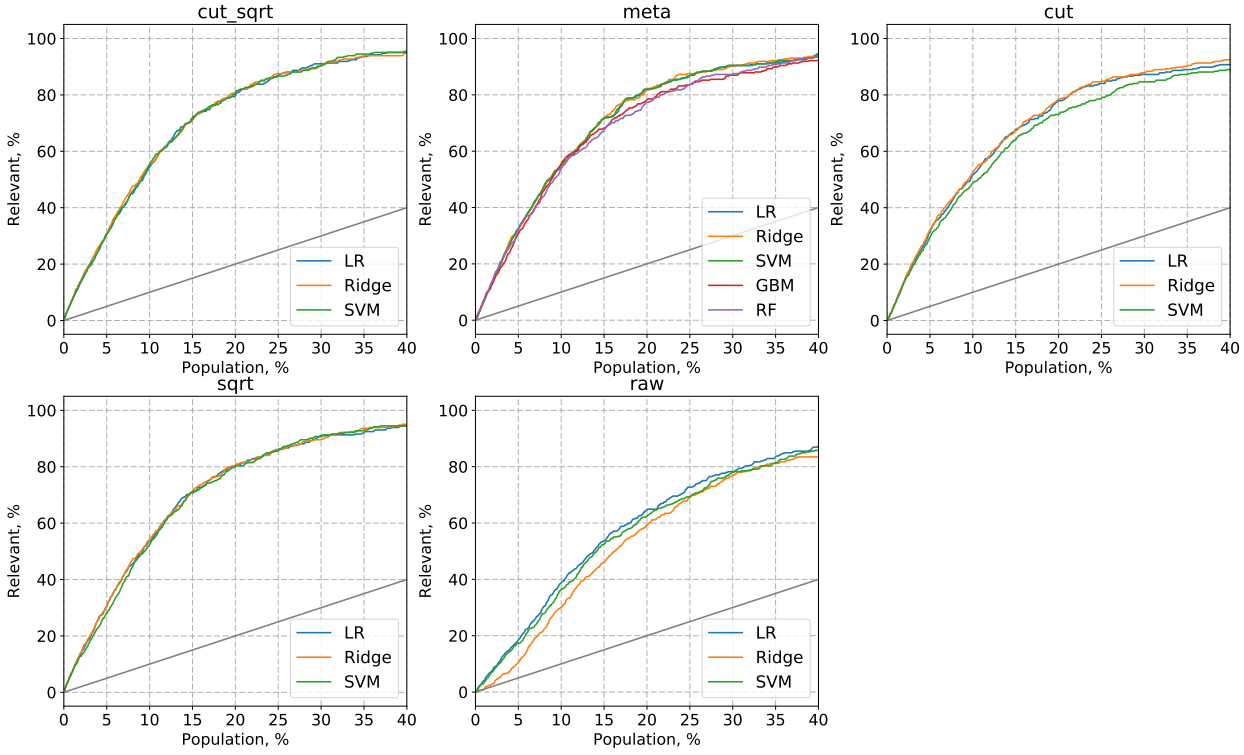


Рис. 11: CG в разбиении по алгоритмам и преобразованию признаков.
Валидация по объединенной выборке Livejournal и IQBuzz.

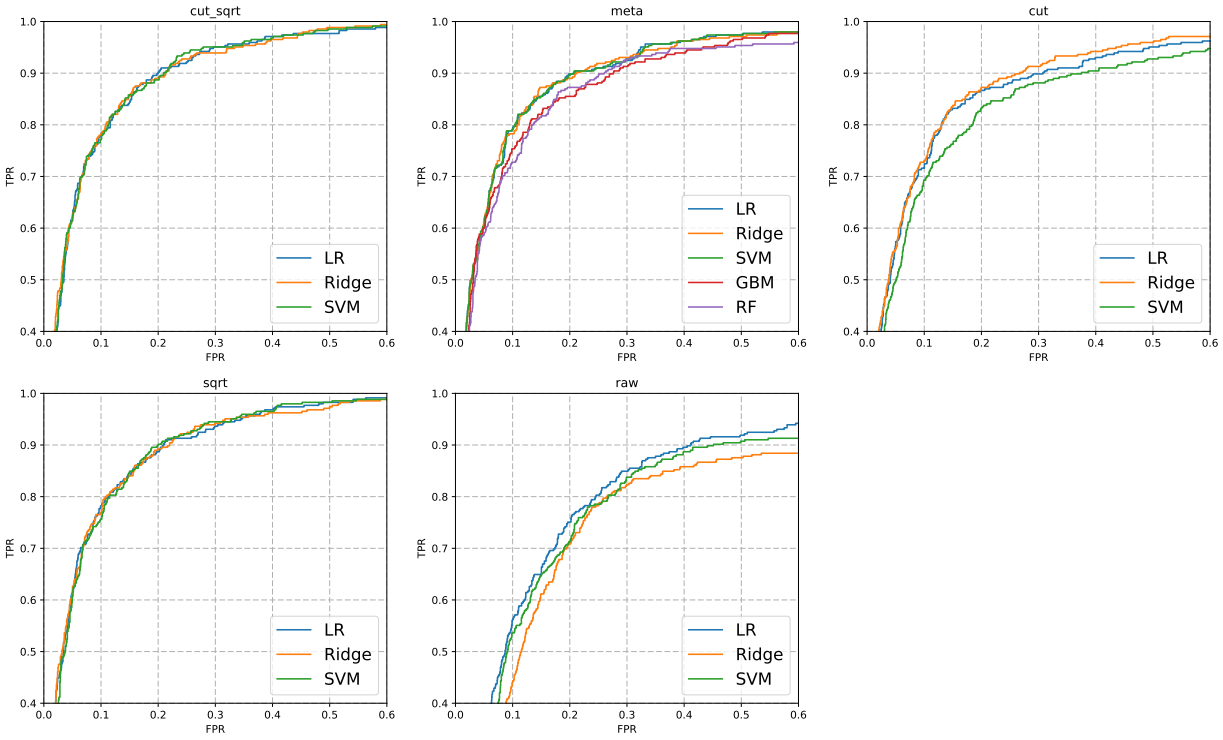


Рис. 12: ROC в разбиении по алгоритмам и преобразованию признаков. Валидация по объединенной выборке Livejournal и IQBuzz.

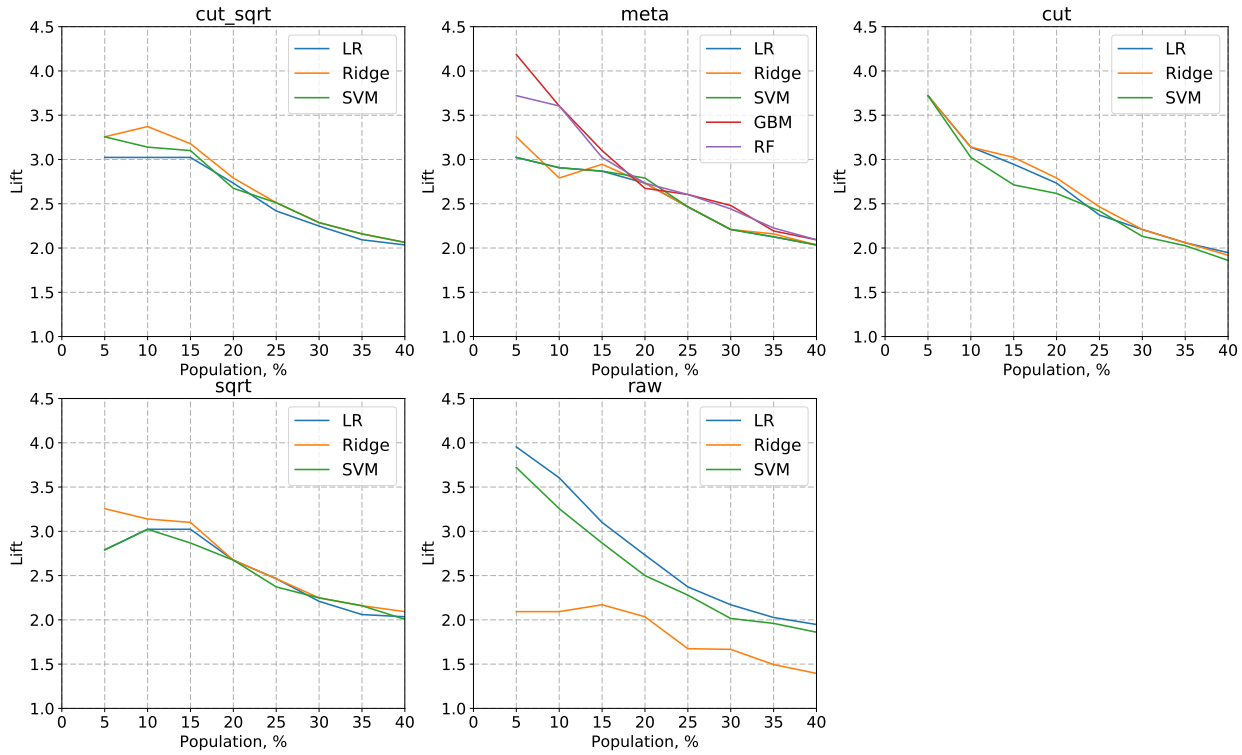


Рис. 13: Lift в разбиении по алгоритмам и преобразованию признаков.
Валидация на выборке IQBuzz, обучение на Livejournal.

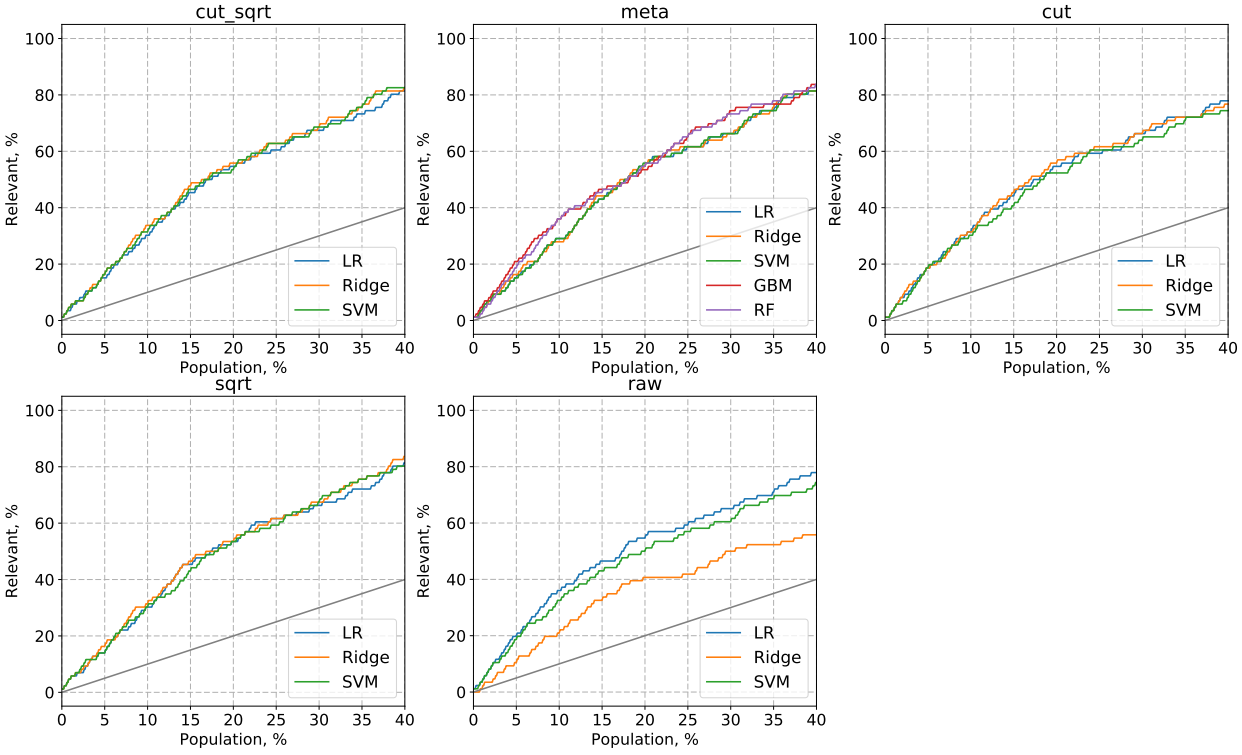


Рис. 14: CG в разбиении по алгоритмам и преобразованию признаков.
Валидация на выборке IQBuzz, обучение на Livejournal.

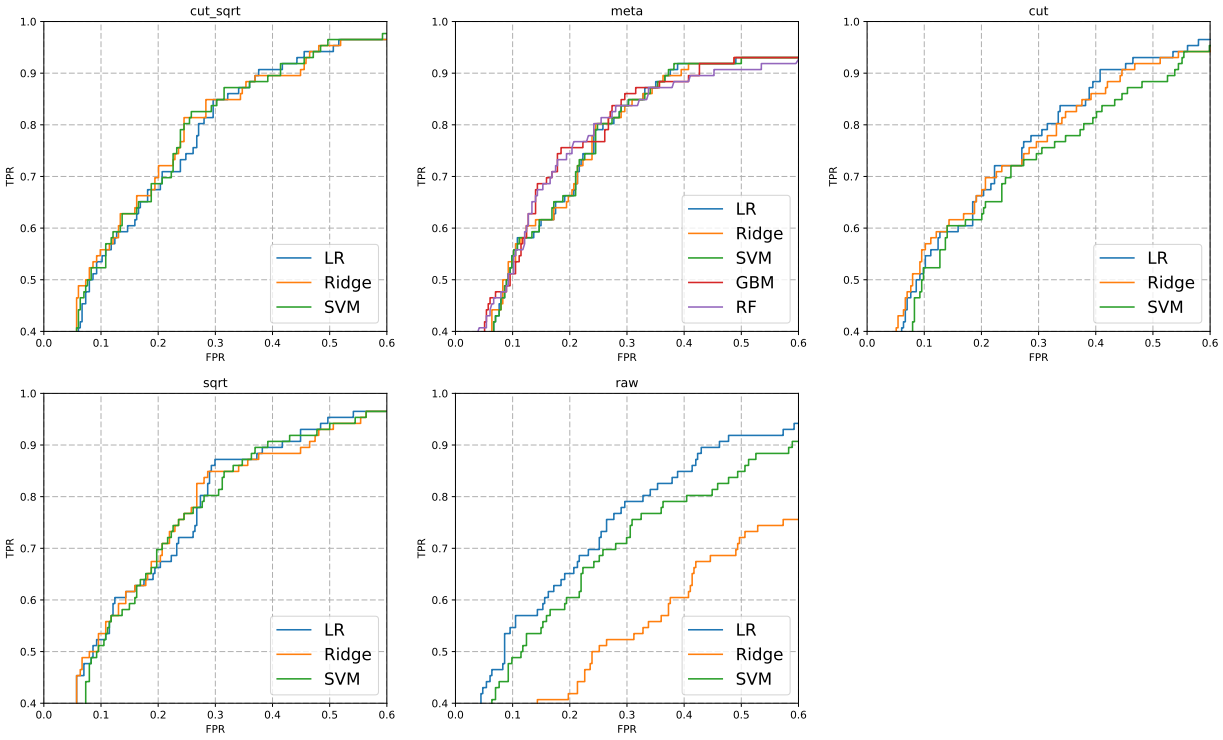


Рис. 15: ROC в разбиении по алгоритмам и преобразованию признаков. Валидация на выборке IQBuzz, обучение на Livejournal.