АВТОМАТИЧЕСКАЯ РЕГИОНАЛЬНАЯ КЛАССИФИКАЦИЯ НА ОСНОВЕ СЛОВАРЯ РЕГИОНАЛЬНОЙ ЛЕКСИКИ: ПРОБНОЕ ИССЛЕДОВАНИЕ

Сорокин А. А. (alexey.sorokin@list.ru)

МГУ им. М. В. Ломоносова, Москва, Россия;
МФТИ, Москва, Россия; РГГУ, Москва, Россия

В данной статье исследуется проблема автоматической региональной классификации на основе подкорпуса ЖЖ (livejournal.com) Генерального Интернет-Корпуса Русского Языка (ГИКРЯ), для этого используется географическая информация, извлечённая из авторских профилей. Поскольку большинство ЖЖ-текстов не демонстрирует достаточно региональных особенностей для надёжной региональной привязки, мы не ставим целью определить регион для всех авторов, однако в случае определения надёжность присваиваемой метки должна быть максимально велика. В качестве признаков используются слова из «Словаря языка русских городов», а в качестве классификатора — наивный Байесовский классификатор, метод опорных векторов и логистическая регрессия. Для оценки уверенности классификатора используется найденная им вероятность класса. В случае 10 удалённых друг от друга регионов точность классификации достигает 97 %, притом региональная метка присваивается 13 % текстов, в то время как для 50 регионов при незначительном падении точности (96 %) полнота падает до 0,5 %.

Ключевые слова: автоматическая классификация, региональная классификация, мультиклассовая классификация, «Словарь языка русских городов», взвешивание признаков

AUTOMATIC REGIONAL CLASSIFICATION USING A DICTIONARY OF REGIONAL LEXICS: A PRELIMINARY STUDY

Sorokin A. A. (alexey.sorokin@list.ru)

Lomonosov Moscow State University, Moscow, Russia;
Moscow Institute of Physics and Technology, Moscow, Russia;
Russian State University of Humanities, Moscow, Russia

Using an automatically collected subcorpus of the Russian segment of livejournal.com, which has annotation for geographic regions of some of the authors, we try to predict the location for the texts which lack such information. Given that the majority of texts in the corpus do not have any regional peculiarities, we try to solve a less ambitious task: to predict regional labels only for a minor fraction of texts, but on such texts our classifier should be accurate. We use different classifiers, such as Naive Bayes, logistic regression and linear SVM, with regional words as features and the predicted probabilities as confidence scores. In case the regions under consideration are located sufficiently apart from each other, the accuracy for regionally specific texts reaches 97 % with 13 % of documents being assigned to some region. For close neighbours the accuracy slightly degrades to 96 %, but the percentage of retrieved documents drops down to 0.5 %.

Keywords: automatic text classification, regional classification, multiclass classification, feature weighting

1. Введение

Для проведения многих социолингвистических и лексикологических исследований, например, исследования региональной вариативности той или иной лексической единицы, полезно и зачастую необходимо наличие в корпусе региональной метатекстовой разметки [3]. В противном случае корпус может оказаться несбалансированным по тому или иному параметру, что может привести к неверной интерпретации результатов корпусных исследований. При автоматическом сборе корпуса с таких сетевых ресурсов, как vkontakte.ru, livejournal.com, blogs.mail.ru и многих других соответствующая информация может быть получена из авторских профилей. Также региональная разметка присутствует и в текстах, извлечённых из прессы, локальных форумов и т. д. Однако в социальных сетях, например, многие авторы не заполняют соответствующее поле или не предоставляют полной или достоверной информации. Например, в используемом в данной статье подкорпусе m.livejournal.com 15 485 авторов указали в своём профиле в поле «регион» значение Russian Federation и 17 306 — Украина, в то время как для второго по распространённости региона России — Петербурга — имеется лишь 12 956 авторов. Более того, значение NA (not available), присваивавшееся в том случае, когда региональную информацию не удавалось извлечь, встречается в 79 077 авторских профилях, что значительно превышает число авторов даже для самого частотного региона (Москвы, 47 709). Таким образом, неатрибутированные локации могут серьёзно повлиять на результаты любого лексикографического или социолингвистического исследования, использующего региональную разметку. Кроме того, сама информация в профилях нуждается в дополнительной проверке. Даже если исключить неизбежные курьёзные локации, такие как Зимбабве (90 авторов) и Гондурас (109 авторов), указанный в профиле регион может указывать как на место фактического проживания, так и на место рождения, учёбы, работы, временного пребывания и т. д.

Таким образом, задача автоматической региональной классификации неизбежно должна быть решена при автоматическом построении Интернет-корпусов. Эту задачу не следует смешивать с задачей геолокации, где требуется указать фактическое местоположение автора текста. Например, если автор вырос в Казахстане, но постоянно проживает в Санкт-Петербурге (и описывает в своих текстах Санкт-Петербург), то с точки зрения геолокации его текстам должна быть приписана метка «Санкт-Петербург», а с точки зрения региональной классификации — Казахстан. Поэтому те признаки, которые при геолокации выходят на первый план (прежде всего, стандартные топонимы), не слишком полезны при решении нашей задачи.

Задача региональной классификации является весьма сложной с алгоритмической точки зрения. Прежде всего, она является мультиклассовой (то есть множество возможных ответов содержит более двух элементов). Это приводит к росту числа настраиваемых параметров алгоритмов, что неизбежно вызывает ухудшение качества классификации. Достаточно указать, что в случае двух классов случайный классификатор будет в среднем правильно классифицировать половину объектов, в то время как для N классов — лишь $1/N$ часть. Кроме того, зачастую при мультиклассовой классификации классы являются несбалансированными (то есть содержат разное число объектов), что приводит к тому, что неправильно обученный алгоритм будет стремиться приписать все объекты к наиболее частотным классам, полностью «забывая» про остальные ([7]).

Указанное свойство мультиклассовой классификации особенно неприемлемо в нашем случае. Действительно, раз мы хотим использовать полученную автоматическую разметку наравне с априорной в будущих исследованиях, и даже уточнять априорные региональные метки с её помощью, то её качество должно быть ничуть не ниже априорной. В качестве разумной оценки требуемой точности можно взять 90%. Разумеется, такая точность в принципе недостижима на любых реальных данных при использовании любого, сколь угодно мощного алгоритма машинного обучения (например, в работе [6], где проводится географическая классификация микроблогов, достигается точность лишь в 24%). Причина состоит в том, что ЖЖ как корпус весьма неоднороден и содержит тексты самой различной природы. У текстов некоторых жанров (например, юридических текстов, энциклопедических статей, кулинарных рецептов и др.) будет вообще отсутствовать выраженная региональная специфика. Встретив такие тексты, наш алгоритм должен отказываться от классификации. Более того, в каком-то смысле слишком высокое качество классификации для таких текстов хуже, чем низкое: ошибка в классификации свидетельствует о том, что алгоритму не удалось найти признаки, сближающие данный текст с другими текстами из того же региона (и таких признаков действительно не должно быть), в то время как правильная классификация свидетельствует либо о переобучении, либо о том, что алгоритм использовал для классификации не региональные, а, например, жанровые, тематические или другие признаки. Априори избежать такой ситуации нельзя: при автоматическом сборе текстов («кроллинге») может случиться так, что распределение текстов по жанрам/тематике/возрасту авторов будет существенно отличаться от региона

к региону. Чтобы устранить данную проблему, потребуется как минимум с высокой точностью определять жанр Интернет-текстов, в то время как само понятие жанра в применении к интернет-текстам зачастую является слишком размытым ([18]).

Исходя из всего сказанного, нашей задачей является не построение алгоритма, дающего неплохую точность для произвольных Интернет-текстов, а получение высокой точности, пусть и на небольшом проценте текстов. Это значит, что нам важна не полнота классификации, а только лишь её точность. При этом алгоритм может и должен отказываться от классификации в тех случаях, когда он недостаточно уверен в правильном ответе. В качестве меры уверенности алгоритма мы используем вероятность выбранного класса. Как известно, наивный байесовский классификатор ([9]) и логистическая регрессия вычисляют данные вероятности в процессе классификации, а для машины опорных векторов ([5]) они могут быть получены за счёт небольшой модификации алгоритма. Насколько известно авторам, подобная задача ранее не ставилась. Более того, практически отсутствуют исследования по автоматической региональной классификации для русскоязычного интернета (можно отметить лишь работу [14], впрочем, достигнутые там результаты весьма невысоки).

В качестве признаков для классификации мы используем слова из «Словаря языка русских городов» ([1], [2], http://community.lingvo.ru/goroda/dictionary.asp). Поскольку блогосфера послужила лишь одним из источников при составлении данного словаря, мы провели предварительное исследование о том, насколько хорошо данные «Словаря языка русских городов» совпадают со статистикой, извлекаемой из корпуса livejournal.com. Наша работа состоит из следующих частей: статистического исследования распределения лексем из «Словаря языка русских городов» по регионам, описания процедуры автоматической классификации с использованием различных алгоритмов классификации и различных методов обработки данных, анализа её результатов и обсуждения дальнейшего применения полученных результатов.

2. Статистический анализ содержания региональной лексики в корпусе m.livejournal.com

На первом этапе исследования мы провели статистический анализ вхождений региональных словоформ в корпус m.livejournal.com. Предполагалось, что словарь региональной лексики не содержит неверно приписанных регионов, хотя может быть существенно неполон. При статистическом анализе мы столкнулись с тем, что локации в словаре не стандартизованы, в результате на предварительном шаге исследования была проведена стандартизация. Большинство стандартных локаций представляют собой регионы Российской Федерации и Украины. Кроме того, список стандартных локаций содержит некоторые регионы Белоруссии и Казахстана, а также страны ближнего зарубежья. К сожалению, множество стандартных локаций не является дизъюнктным и содержит некоторые надрегиональные локации, такие как Западная Сибирь,

Урал, Сибирский ФО, Дальний Восток, а также Украина и Белоруссия. В результате процедура проверки соответствий между словарём и корпусом не сводится к простому соответствию: корпусные локации для словарных слов могут быть не только положительными и отрицательными, но и нейтральными (как например, корпусная локация «Урал» для словарной локации «Челябинская область»). Также нейтральными считались все локации из корпуса, которые не удавалось стандартизовать (например, NA).

Статистика вхождений была подсчитана на корпусе m.livejournal.com. Он представляет собой подкорпус корпуса ГИКРЯ и после лемматизации и морфологической разметки имеет размер 183 ГБ. Мы использовали список из 693 региональных слов, представляющих собой наиболее надёжно атрибутированные регионализмы. Данный список в не включает в себя те лексемы, региональность которых заключается в узусе, а не в самом факте употребления (например, были исключены «башня» и «свечка» в значении «высокий и узкий многоэтажный дом»). Для отобранных слов в ходе предыдущих корпусных исследований Р. Идрисовым была построена полная парадигма, что сняло необходимость прибегать при поиске к использованию морфологических анализаторов. Для того, чтобы исключить омонимию словоформ из словаря с формами несловарных слов, использовался следующий метод: каждая словоформа в словаре подавалась на вход анализатору mystem; в случае, если анализатор возвращал лемму, не совпадающую со словарной, данная словоформа не учитывалась при расчётах. Например, слово «курам» в выражении «курам на смех» не считалось формой лексемы «кура», поскольку для него существует несловарная лемма «курица». В сочетании с предварительным устранением многозначных лексем это позволило осуществлять поиск простой проверкой слов на совпадение.

Подсчёт статистик вёлся как по числу вхождений слов в корпус, так и по числу авторов, употребляющих данную региональную лексему. Поскольку статистика «по авторам» является гораздо менее чувствительной к выбросам и побочным факторам, то в дальнейшем исследовании используется именно она. Таким образом, за один «документ» принималась вся совокупность текстов данного автора. При этом мы не учитывали «документы» суммарной длиной менее 5000 слов. Ниже в таблице приведена статистика по 10 наиболее частотным региональным лексемам и по 10 наиболее частотным регионам (мы исключили из таблицы составные надрегионы России, такие как Урал и Дальний Восток).

Таблица 1. Статистика вхождений по региональным лексемам

Лексема	Общее число вхождений	Положительные вхождения	Отрицательные вхождения	Нейтральные вхождения
сотовый	37 229	788	16 490	19 451
греча	8 018	759	3 260	3 999
мобилка	8 496	1 343	3 215	3 938
занос	7 496	1 584	2 348	3 564
обменник	5 636	107	2 688	2 841

Лексема	Общее число вхождений	Положительные вхождения	Отрицательные вхождения	Нейтральные вхождения
Новосиб	3666	727	1329	1610
шаверма	3617	866	1109	1642
поребрик	4049	1046	1107	1896
догонялки	4527	55	2304	3628
ботсад	2648	1216	323	1109

Таблица 2. Статистика вхождений по регионам

Регион	Общее число авторов	Положительные вхождения	Отрицательные вхождения	Нейтральные вхождения
Москва	47709	9	472	0
Украина	17306	77	268	42
Петербург	12956	24	320	0
Московская область	3533	11	210	0
Свердловская область	2274	54	151	0
Новосибирская область	2060	64	150	0
Самарская область	1518	14	141	0
Белоруссия	1493	18	138	3
Краснодарский край	1180	39	130	0
Ростовская область	1177	39	120	0

Разумеется, из такой статистики ещё нельзя делать существенные отрицательные выводы. Гораздо более показательными, чем абсолютные частоты, являются относительные величины (например, среднее количество авторов, употребивших данную лексему, на 100 авторских профилей). Для каждой лексемы w мы измерили следующий показатель: пусть в словаре для неё имеется t регионов, обозначим через $c_+(w)$ среднюю долю авторов в данных регионах, употребляющих данную лексему (усреднение ведётся по регионам, а не по авторам). Аналогично обозначим через $c_-(w)$ среднюю долю употребления данной лексемы в тех t отрицательных регионах, где она встречается чаще всего. Ниже в таблицах приведено распределение значений величины $r(w)$ по интервалам для 598 региональных лексем, встречающихся в корпусе. Чтобы дать более полную картину и исключить случайные отклонения, мы приводим две таблицы, в первой из которых учитываются все регионы, а во второй — только те, где данная лексема встречается не меньше 5 раз.

Таблица 3. Распределение отношения ***r*** средней частоты встречаемости лексемы в положительных и отрицательных регионах (с учётом и без учёта регионов, содержащих менее 5 вхождений данной лексемы

r	число регионов
$r \geq 5{,}0$	133
$5{,}0 \geq r \geq 2{,}0$	65
$2{,}0 \geq r \geq 1{,}0$	70
$1{,}0 \geq r \geq 0{,}5$	72
$0{,}5 \geq r \geq 0{,}0$	153
$r = 0{,}0$	105

r	число регионов
$r \geq 5{,}0$	171
$5{,}0 \geq r \geq 2{,}0$	26
$2{,}0 \geq r \geq 1{,}0$	24
$1{,}0 \geq r \geq 0{,}5$	19
$0{,}5 \geq r \geq 0{,}0$	19
$r = 0{,}0$	439

Таким образом, 105 из 698 лексем вообще не встретилось в тех регионах, в которых они употребляются согласно словарю, а 439 были употреблены менее, чем 5 авторами. Разумеется, здесь следует сделать поправку на неполноту корпуса для некоторых регионов, однако в любом случае это говорит о существенной неполноте регионального словаря. Заметим, что если вычислить данную статистику лишь для 100 наиболее частотных лемм (с выбрасыванием регионов, где имеется менее 5 авторов, употребивших данную лексему), то для 31 из них отношение *r* будет больше 5,0, а ещё для 14 — больше 2,0, что уже гораздо более приемлемо. Стоит отметить, что без удаления регионов с недостаточным числом авторов неравенство $r > 1$ выполняется лишь для 31 региона, то есть данная процедура существенно улучшает надёжность данных. Таким образом, аномально большое число несовпадений в большей степени объясняется несбалансированностью корпуса и несовершенством методики, чем недостатками словаря. В нашем исследовании мы не ставим своей целью уточнить словарь[1], для нас существенно, что его данные не могут быть использованы для автоматической региональной привязки текста в данном корпусе.

Однако мы можем использовать саму корпусную статистику, взяв из словаря лишь список лемм. Но при этом следует проверить, что данная статистика обладает хорошей предсказательной способностью: в частности, что с точки зрения встречаемости региональных слов близкие регионы действительно близки, а далёкие — далеки. Для этого мы рассмотрели задачу кластеризации регионов на основе близости соответствующих им эмпирических распределений на множестве региональных слов. В качестве признаков мы взяли частоты слов в регионе, вычисленные «по авторам», после чего нормировали их на общее количество авторов для данного регионам. Таким образом, каждый регион оказался представлен 698-мерным вектором. В качестве функции расстояния

[1] По сообщению В. Е. Беликова, «Словарь языка русских городов» должен подвергнуться существенной модификации, прежде чем он будет использован в автоматической региональной классификации. Например, он должен быть очищен от орфографических регионализмов, таких как «растягай», так как в этом случае невозможно отличить региональную вариативность написания от простой орфографической ошибки

использовалась обычная евклидова метрика без взвешивания признаков. Мы использовали пакет cluster.hierarchy библиотеки scipy ([8]), написанной на языке Python, для определения числа кластеров использовался метод Сальвадора (в просторечии «метод колена», [17]). Было обнаружено, что данные содержат явные сгущения, автоматически были выделены следующие 6 кластеров (см. таблицу 4).

Таблица 4. Распределение регионов по кластерам

Название кластера	**Регионы**
Украина (с Крымом)	
Казахстан	Казахстан, Алма-Ата
Дальний Восток	Амурская область, Дальний Восток, Забайкальский край, Приморский край, Сахалинская область, Хабаровский край, Хакасия
Урал	Кировская область, Пермский край, Свердловская область, Тюменская область, Удмуртия, Урал, Челябинская область
Сибирь	Сибирский ФО, Алтайский край, Западная Сибирь, Иркутская область, Кемеровская область, Красноярский край, Новосибирская область, Омская область, Томская область
Остальное	европейская часть России, Прибалтика, Молдавия, Закавказье, Узбекистан

Ниже мы приводим более детальную дендрограмму (отделение приведено по самому нижнему разрезу, построенному по методу Сальвадора). Видно, что в один кластер действительно попадают географически близкие регионы (за исключением отнесённой к Дальнему Востоку Хакасии). Интересно отметить, что самыми близкими с точки зрения выбранной меры оказываются родо-видовые пары регионов Алма-Ата-Казахстан, Киев-Украина и Свердловская область-Урал. Таким образом, либо большинство авторов, указавших в качестве региона Украину, в действительности проживают в Киеве или Киевской области, либо указанные регионы служат своеобразными центроидами для полученных кластеров.

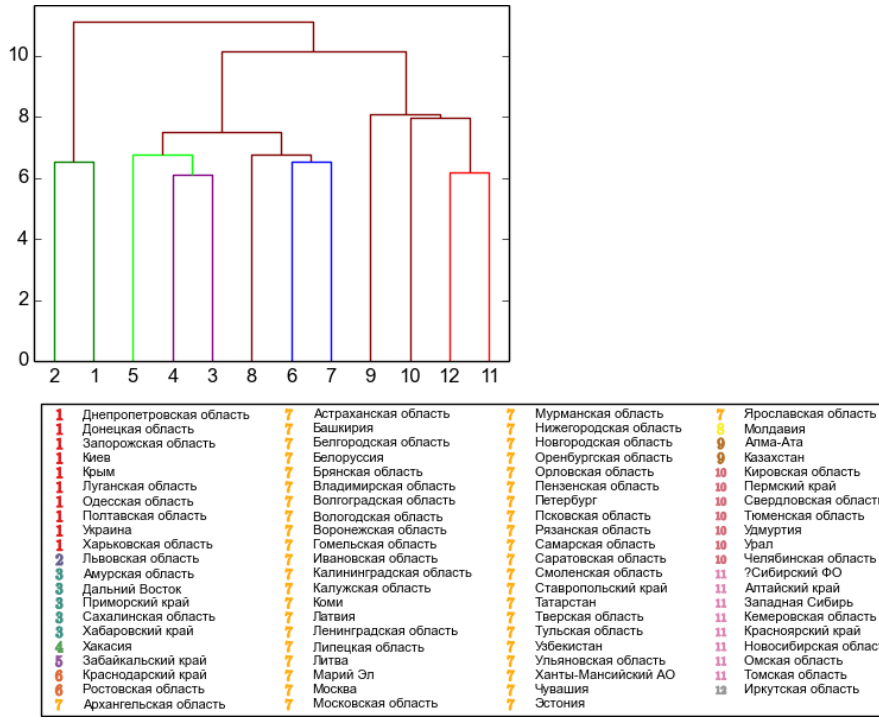

Рисунок 1. Дендрограмма для кластеризации регионов по частотам региональных слов

3. Автоматическая региональная классификация: описание данных и алгоритмов

Данный раздел посвящен описания алгоритмов, использованных для автоматической региональной классификации. Базовой моделью в текстовой классификации является модель «мешка слов»; в ней учитывается только число вхождений того или иного термина в текст, но не порядок и взаимное расположение таких вхождений. Обозначим через N_{ij} значение j-го признака для i-го объекта. Существует две базовых разновидности модели «мешка слов»: многомерная, в которой в качестве N_{ij} берётся индикатор вхождения j-го термина и мультиномиальная, где N_{ij} равно числу таких вхождений. Первоначально данные термины были введены для наивного байесовского классификатора в работе [9], однако они характеризуют не столько сам алгоритм классификации, сколько метод извлечения данных из текста. Для большинства задач мультиномиальная модель более предпочтительна, т. к. она более полно использует информацию, содержащуюся в тексте и менее чувствительна к зависимости признаков. Многомерная модель применяется в основном в случае, когда имеется небольшое

количество признаков, являющихся сильными положительными предикторами[2]. Отметим, что такая ситуация как раз наблюдается в нашей задаче. Также мы использовали лог-мультиномиальную модель ([16]), получающуюся из мультиномиальной путём трансформации $N_{ij} \mapsto \log_2(1 + N_{ij})$. Её преимущество состоит в том, что она более точно отражает распределение частоты слова в тексте, чем мультиномиальная модель. Напомним, что в нашей задаче документам соответствуют все тексты одного автора, а вхождениям — употребления термина в тексте. Таким образом, в многомерной в качестве признаков мы брали индикаторы употребления региональных лексем данным автором, а в мультиномиальной — число текстов, в которых употреблялась данная лексема.

Считается, что в большинстве задач наивный байесовский классификатор, особенно в многомерной версии, проигрывает другим линейным классификаторам (таким как логистическая регрессия и машина опорных векторов). Причина этого лежит в нарушении предположения о независимости признаков, лежащего в основе байесовской модели. В нашем случае в первом приближении вхождения разных региональных слов в текст можно считать независимыми, поэтому данным аргументом можно пренебречь. Кроме того, было показано, что после надлежащего взвешивания и отбора признаков в отдельных задачах наивный байесовский классификатор может показывать сравнимые с более сложными моделями результаты ([16]). В связи с этим мы решили проверить все три классификатора (наивный байесовский, логистическую регрессию и машину опорных векторов) в сочетании с различными методиками отбора признаков. Значимость признаков вычислялась отдельно для каждого класса, таким образом, каждому признаку w_j сопоставлялся набор весов w_{ij}. Использовалось четыре методики взвешивания признаков:

1. Логарифмическое отношение вероятностей (log odds, [12]):

$$w_{ij} = \log \frac{P(w_j|c_i)(1 - P(w_j|\overline{c_i}))}{(1 - P(w_j|c_i))P(w_j|\overline{c_i})}$$

2. Взаимная информация (information gain):

$$w_{ij} = H(w_j) - H(w_j|c_i)$$

3. Вероятность класса (ambiguity measure, [11]): $w_{ij} = P(c_i|w_j)$.
4. Вес признака (feature weight, [13]): w_{ij} равен соответствующему весу в линейной модели.

Здесь c_i — индикатор принадлежности документа классу (то есть региону), а w_j — индикатор вхождения термина в текст. В случае многомерной модели вероятности считались «по авторам», а в случае мультиномиальной — «по текстам». Отбор признаков позволяет исключить из модели те лексемы, которые встречаются в слишком большом числе регионов (например, «сотовый», «греча» или «кура») и потому не являются хорошими предикторами. После этого мы отбирали фиксированное количество признаков с наибольшими весами

[2] Данный вопрос детально исследовался в [9]

по следующей схеме ([7]): мы по очереди рассматривали все регионы и для каждого из них выбирали признак с наибольшим весом из ещё не отобранных. Так делалось до тех пор, пока число признаков не достигало требуемого количества. Данный метод позволяет отобрать небольшое количество признаков, являющихся хорошими предикторами для плохо предсказываемых классов, в то время как при использовании всех признаков их влияние перевешивается остальными.

Зачастую причиной низких результатов классификации является несбалансированность исходных данных или их недостаточность, а также близость некоторых классов в пространстве признаков. В связи с этим мы проводили исследования только для 50 наиболее частотных регионов, упорядоченных по степени удалённости друг от друга (то есть вначале в списке регионов идут два самых удалённых друг от друга, потом самый удалённый от них и т. д.). Расстояние между регионами считалось так же, как при их кластеризации в разделе 2. При этом мы удалили из списка неэлементарные регионы (Украина, Дальний Восток, Урал, Сибирский ФО), а также Москву (из-за отсутствия региональной специфики). После этого в выборке осталось 43 региона, список которых приведен в таблице ниже.

Таблица 5. Распределение регионов по группам

1–10	11–20	21–40
Иркутская область	Одесская область	Воронежская область
Калининградская область	Петербург	Пермский край
Оренбургская область	Хабаровский край	Томская область
Тверская область	Удмуртия	Молдавия
Новосибирская область	Латвия	Тюменская область
Алма-Ата	Красноярский край	Челябинская область
Львовская область	Киев	Волгоградская область
Свердловская область	Белоруссия	Кемеровская область
Крым	Башкирия	Эстония
Ульяновская область	Саратовская область	Омская область
		Ростовская область
		Алтайский край
		Донецкая область
		Приморский край
		Харьковская область
		Ярославская область
		Алтайский край
		Московская область
		Татарстан
		Днепропетровская область
		Самарская область
		Краснодарский край
		Казахстан

Для каждого региона авторы упорядочивались по количеству употреблений региональной лексики, после чего отбиралось фиксированное количество лидирующих авторов, одинаковое для каждого класса. Чтобы исследовать влияние близости классов на качество классификации, в различных экспериментах выбиралось разное число наиболее удалённых друг от друга регионов. Описание результатов эксперимента приведено в следующем разделе.

4. Анализ результатов

Ещё раз опишем те параметры, которые варьировались при классификации:

1. Модель представления данных (многомерная, мультиномиальная, лог-мультиномиальная).
2. Число классифицируемых регионов (10, 20, 44).
3. Число авторов для каждого региона (100, 500).
4. Метод отбора признаков (логарифмическое отношение вероятностей (ЛОВ), информация, вероятность класса (ВК), вес признаков (ВП), отсутствие отбора (ОО)).
5. Число отбираемых признаков (100, 200).
6. Алгоритм классификации (наивный байесовский, логистическая регрессия, машина опорных векторов).

Поскольку нас интересует только точность классификации для объектов, имеющих высокую вероятность отнесения к тому или иному классу, то в качестве меры качества мы использовали точность классификации для документов, которым классификатор сопоставлял вероятность более 0.9. При этом в качестве общей точности бралось среднее значение данной величины по всем регионам. При классификации мы случайным образом разбивали выборку на обучающую и контрольную в отношении 4/1 (то есть 80% объектов попадало в обучение и 20% в контроль), при этом распределение вероятностей классов не отличалось между обучающей и контрольной выборкой. Мы повторяли данное разбиение 10 раз, после чего результаты усреднялись.

Мы взяли реализацию логистической регрессии и машины опорных векторов из пакета scikit-learn ([15]), написанного на языке Python. Данная имплементация основана на библиотеке LIBSVM ([4]) Реализация наивного байесовского классификатора была написана самостоятельно, при этом мы не учитывали априорные вероятности классов. Для предобработки данных также использовались средства пакета scikit-learn.

Ниже мы приводим результаты экспериментов. Для каждого алгоритма мы указываем наилучшую схему отбора признаков. Результаты для мультиномиальной модели не приводятся, поскольку она существенна уступала двум другим вариантам. Данные собраны в 3 таблицы, для 10, 20 и 44 регионов. Сразу за точностью классификации для объектов, вероятность отнесения которых к нужному классу превысила 0.9, мы приводим долю таких объектов. Для сравнения мы также приводим общую точность классификации. В том случае, если для данной пары (алгоритм, модель) в каком-то классе не оказывалось

объектов с вероятностью отнесения к данному классу выше порога 0.9, мы приводили значение для следующего порога вероятности (0.75). Данные случаи специально помечены в таблице.

Таблица 6. Результаты классификации для 10 регионов

Число авторов	Модель	Наивный байесовский классификатор	Логистическая регрессия	Машина опорных векторов
100	Многомерная	ЛОВ 200	ЛОВ 100	ИНФ 100
		97.4 26.0 73.4	99.0 7.5 76.4	100.0 9.9 64.1
100	Логмультиномиальная	КВ 100	ОО	ВП 100
		93.6 44.1 70.9	100.0 8.4 66.0	100.0 9.7 61.3
500	Многомерная	ЛОВ 100	КВ 200	ИНФ 200
		96.7 13.3 76.5	99.3 2.5 54.1	97.0 7.0 53.8
500	Логмультиномиальная	ВП 100	ВП 200	ВП 100
		92.7 26.1 62.2	97.9 6.3 54.5	93.7 4.0 65.3

Таким образом, в случае наивного байесовского классификатора многомерная модель показывает стабильно более высокие результаты, чем логмультиномиальная. При этом в качестве весовой функции для признаков используется логарифмическое отношение вероятностей, полезность которого для наивного байесовского классификатора неоднократно отмечалась ранее ([12]). Логистическая регрессия и машина опорных векторов дают несколько более высокие результаты, однако при этом отбирается значительно меньшее число объектов. В случае большего числа классов это может привести к тому, что для некоторых регионов вообще ни один текст не получит достаточно высокую вероятность. В целом же все три классификатора демонстрируют сравнимые высокие результаты.

Таблица 7. Результаты классификации для 20 регионов

Число авторов	Модель	Наивный байесовский классификатор	Логистическая регрессия	Машина опорных векторов
100	Многомерная	ЛОВ 100	ОО	ВП 100
		100.0 3.1 58.3	**100.0 1.4 49.8**	96.7 2.1 45.8
100	Логмультиномиальная	ЛОВ 100	ВК 100	ОО
		93.4 9.5 53.7	93.8 4.2 54.0	100.0 2.7 46.6
500	Многомерная	ЛОВ 100	ВК 200	ИНФ 100
		98.4 1.7 65.1	**92.4 2.2 45.3**	93.8 1.9 49.8
500	Логмультиномиальная	ВП 100	ЛОВ 200	ИНФ 100
		82.5 6.3 60.0	92.1 2.7 65.2	89.6 3.0 49.4

При переходе от 10 регионов к 20 качество классификации существенно не ухудшается. При этом оптимальным вариантом в случае 500 текстов с большим преимуществом оказывается многомерный байесовский классификатор с логарифмическим отношением вероятности в качестве функции весов. Более того, при данных параметрах модели наилучшим оказывается и среднее качество классификации для всех объектов.

Таблица 8. Результаты классификации для 44 регионов

Число авторов	**Модель**	**Наивный байесовский классификатор**	**Логистическая регрессия**	**Машина опорных векторов**
100	Многомерная	ЛОВ 200	ВК 100	ВК 200
		87.7 2.0 36.0	**92.2 2.8 40.3**	**98.3 0.9 29.5**
100	Логмультиномиальная	ЛОВ 100	ВП 200	ОО
		85.5 3.7 33.6	**96.7 0.8 31.3**	96.9 0.8 27.0
500	Многомерная	ЛОВ 200	ВП 200	ВК 200
		96.8 0.5 47.4	**97.5 0.1 28.7**	87.2 0.2 31.2
500	Логмультиномиальная	ЛОВ 100	ЛОВ 200	ВК 100
		88.4 1.5 39.0	**86.4 0.4 42.3**	78.5 0.3 24.8

В случае 44 классов сохраняется преимущество многомерной модели над мультиномиальной. При этом для 500 объектов наивный байесовский классификатор даёт почти наилучший результат, лишь 0.7% уступая логистической регрессии. При этом доля объектов, классифицируемых с высокой вероятностью, для него существенно выше. Следует отметить, что при большом числе классов переход от 100 авторов к 500 не ухудшает качества классификации (что могло бы произойти из-за участия в классификации текстов с более низким содержанием региональной лексики), а напротив, увеличивает (за счёт увеличения обучающей выборки). Кроме того, следует отметить, что наивный байесовский классификатор стабильно показывает наилучшие результаты при отборе признаков с помощью логарифмического отношения вероятностей. Более того, использование данного метода отбора признаков приводит к значимому росту точности классификации по сравнению с остальными методами (соответствующие результаты не приведены из-за недостатка места). Интересно отметить, что для других алгоритмов классификации соотношение между разными методами отбора признаков меняется и более предпочтительным оказывается метод вероятности класса.

5. Обсуждение и применение результатов

Результаты предыдущего раздела показывают, что автоматические методы классификации (и прежде всего многомерный байесовский классификатор с отбором признаков с помощью логарифмического отношения вероятностей)

могут успешно применяться для региональной разметки. Это является основным положительным результатом, поскольку до проведения данного исследования возможность автоматической классификации при неточном словаре была неочевидна. Кроме того, столь высокое качество классификации показывает, что и априорная разметка, на основе которой она проводилась, была достаточно точной. Следующим этапом эксперимента могла бы быть автоматическая классификация текстов с корпусной локацией NA на основе наивного байесовского классификатора, обученного на текстах, содержащих данную информацию.

Однако следует отметить некоторые отрицательные моменты: прежде всего это невысокий процент присваиваемых меток: определяются локации лишь 1–2% новых текстов, что достаточно мало. При уменьшении порога вероятности качество классификации падает незначительно (до 88%), однако и процент извлекаемых текстов не увеличивается. Существенные различия между случаем 10 и 44 регионов показывают, что наивный байесовский классификатор можно использовать для грубого определения примерной региональной метки, после чего уточнение производить с помощью более сложных моделей. Также улучшения качества классификации и более надёжного разделения классов можно добиться за счёт расширения множества региональных слов или уточнения априорных локаций в словаре. К сожалению, результаты региональной классификации плохо переносимы на новый корпус: при попытке применить ту же модель к корпусу m.vk.com оказалось, что тексты оттуда содержат гораздо меньше региональных слов (это объясняется значительно меньшими значениями количества слов на одного автора в корпусе и длины одного текста), что приводит к существенному изменению распределений частот. По-видимому, результаты могут оказаться применимы лишь к текстам примерно той же средней длины, что и в корпусе m.livejournal.com

Другим способом увеличения полноты классификатора может служить усложнение вероятностной модели. Наиболее естественным является введение дополнительного класса, соответствующего текстам без региональной специфики и использования смеси вероятностных распределений ([10]).

В любом случае, построенный в данной работе классификатор может быть успешно применён как минимум для грубого определения региональной метки. Кроме того, результаты данного исследования показывают существенное влияние методов отбора признаков и способов представления данных на результаты классификации. Это может оказаться полезным при выборе алгоритма в будущих исследованиях по региональной классификации.

Благодарности

Автор благодарит Идриса Юсупова и Николая Копылова за помощь в компьютерной обработке данных, В. Е. Беликова за помощь в анализе и интерпретации полученных результатов, а также С. А. Шарова и В. П. Селегея за ценные обсуждения.

References

1. *Belikov V.* (2006). The examples for the dictionary of the varieties of urban Russian and the WWW [Slovar' "Yazyki russkikh gorodov": podbor primerov i Internet], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2006" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2006"], Bekasovo, pp. 57–60.
2. *Belikov V.* (2008). Urban language: materials for the vocabulary of literary lexis [Yazyki gorodov: materialy k slovariu literaturnoj leksiki], ABBYY, Moscow.
3. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013). Corpus as language: from scalability to register variation [Korpus kak yazyk: ot masshtabiruemosti k differentsial'noj polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], Bekasovo, pp. 84–96.
4. *Chang C.-C. an Lin C.-I.* (2011). LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology, Vol. 2, No. 3, pp. 1–27.
5. *Cortes C. and Vapnik V.* (1995). Support-vector networks, Machine learning, Vol. 20, No. 3, pp. 273–297.
6. *Eisenstein J. et al* (2010). A latent variable model for geographic lexical variation. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, ACL, 2010, pp. 1277–1287.
7. *Forman G.* (2004). A pitfall and solution in multi-class feature selection for text classification. In: Proceedings of the 21st international conference on Machine learning, ACM, 2004, p. 38.
8. *Jones E., Oliphant T., Peterson P. et al.* (2001–). Scipy: Open source scientific tools for Python, available at scipy.org
9. *McCallum A., Nigam K. et al.* (1998). A comparison of event models for naive bayes text classification. In: Proceedings of AAAI-98 workshop on learning for text categorization, Vol. 752, pp. 41–48.
10. *McCallum A.* (1999) .Multi-label text classification with a mixture model trained by EM. In: Proceedings of AAAI-99 Workshop on Text Learning, pp. 1–7.
11. *Mengle S. and Goharian N.* (2009). Ambiguity measure feature-selection algorithms. Journal of the American Society for Information Science and Technology, Vol. 60., No. 5, pp. 1037–1050.
12. *Mladenic D., Grobelnik M.* (1999). Feature selection for unbalanced class distribution and naive bayes, ICML., Vol. 99. pp. 258–267.
13. *Mladenic D. et al.* (2004). Feature selection using linear classifier weights: interaction with classification models, In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval, Sheffield, pp. 234–241.
14. *Morozov E. and Bogdanova D.* (2013). Detecting region by Livejournal data [Opredelenie regiona po dannym ghivogo ghurnala], available at http://www.science-education.ru/pdf/2013/4/232.pdf.

15. *Pedregosa F. et al.* (2011). Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, Vol. 12, pp. 2825–2830.
16. *Rennie J. D.* (2003). Tackling the poor assumptions of naive bayes text classifiers. In: Proceedings of ICML, Washington DC, Vol. 3, pp. 616–623.
17. *Salvador S., Chan P.* (2004). Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, In: Tools with Artificial Intelligence, 2004, Proceedings of ICTAI 2004, pp. 576–584.
18. *Sorokin A., Katinskaya A. and Sharoff S.* (2014). Associating symptoms with syndromes: reliable genre annotation for a large Russian webcorpus. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], Bekasovo, pp. 646–659, http://www.dialog-21.ru/digests/dialog2014/materials/pdf/SorokinAKatinskayaASharoffS.pdf