УДК/UDC 81'32+81'33

Исследование методов классификации текстов для несбалансированных данных

Садов Михаил Александрович
бакалавр фундаментальной и прикладной лингвистики, студент магистратуры по направлению «Компьютерная лингвистика» факультета гуманитарных наук
Национальный исследовательский университет Высшая школа экономики
г. Москва, Россия
e-mail: juneau52@mail.ru

Аннотация

В ходе данного исследования были проанализированы методы классификации для несбалансированных данных, а также определён и реализован наиболее эффективный из них с учётом текстовой специфики русского языка. Было установлено, что при работе с несбалансированными данными эффективнее работать с изменением порога решений и весами ошибок разного рода, чем пытаться изменить распределение обучающего набора данных случайным образом или синтетически. В связи с этим наиболее эффективным подходом был признан метод с распределением весов. Для эмпирической проверки теоретических выводов была выбрана задача бинарной классификации высказываний на русском языке в зависимости от того, корректны ли они или содержат какую-либо оскорбительную информацию. В результате анализа результатов было выяснено, что использование методов с распределением весов не помогло улучшить результат классификации текстов по сравнению с обучением модели на умеренно несбалансированных данных для данной задачи. Однако после проведения подобного испытания на сильно несбалансированных данных было установлено, что использование взвешенных моделей действительно приводит к улучшению результата классификации по сравнению с обучением на несбалансированных данных.

Ключевые слова: обработка естественного языка, компьютерная лингвистика, прикладная лингвистика, машинное обучение, несбалансированные данные.

The Research of Methods for Texts Classification for Unbalanced Data

Sadov Mikhail Aleksandrovich
Bachelor of fundamental and applied linguistics, student of Computational Linguistics Master's programme at the Faculty of Humanities
National Research University Higher School of Economics
Moscow, Russia
e-mail: juneau52@mail.ru

Abstract

This research deals with analysing methods for texts classification for unbalanced data. The most effective of these methods was defined and implemented taking into account the peculiarities of the Russian language. It was found that the most effective way to work

with unbalanced data is to change the output threshold or classes weights rather than changing the distribution of the training data set randomly or synthetically. Thus, it was defined that the most effective approach is to use cost-sensitive methods. In order to verify the theoretical conclusions empirically, the task of classifying statements in Russian was selected depending on whether they are correct or contain any offensive content. As a result, it was found that the usage of cost-sensitive methods did not help to improve the results of the texts classification compared to training a model on moderately unbalanced data for this task. However, after conducting such an experiment using strongly skewed data, it was found that working with weighted models indeed leads to an improved classification result.

Key words: natural language processing, computational linguistics, applied linguistics, machine learning, unbalanced data.

Введение. Для успешного решения задач автоматической обработки естественного языка очень важно иметь возможность правильно классифицировать текстовые данные в соответствии с проблемой, стоящей перед человеком. Примером подобной задачи является фильтрация спама в электронных почтовых ящиках (определение, принадлежит ли поступившее владельцу письмо к классу «спам» или нет).

В качестве методов, используемых для установления принадлежности объекта к тому или иному классу, чаще всего используют машинное обучение с учителем. Основной идеей этого подхода является индуктивный вывод функции на основе размеченных данных для обучения [1]. Это означает, что успешность применения алгоритма машинного обучения с учителем во многом зависит от той выборки объектов, на основе которых «обучается» программа. Большинство подобных алгоритмов требуют от исследователя включения сопоставимого количества примеров для каждого из классов, однако зачастую сделать сбалансированные наборы данных не представляется возможным в связи с рядом факторов. Ключевыми из них являются специфика целевой области (балансировка данных может понизить показатель их репрезентативности — «меры возможности восстановить, воспроизвести представление о целом по его части или мера возможности распространить представление о

части на включающее эту часть целое» [2]) и разная цена ошибок первого и второго рода при классификации.

Вследствие этого возникает проблема обучения модели на несбалансированных данных: в соответствии с базовыми предположениями, заключёнными в большинство алгоритмов, целью обучения является максимизация доли правильных решений по отношению ко всем принятым решениям, а данные для обучения и генеральная совокупность подчиняются одному и тому же распределению [3]. Однако учёт данных предположений и несбалансированности выборки приводит к тому, что модель оказывается неспособна классифицировать данные согласно алгоритму лучше, чем тривиальная модель, полностью игнорирующая менее представленный класс и маркирующая все объекты для классификации как принадлежащие к доминирующему классу.

Другой вариант развития событий — это создание слишком сложной модели, включающей большое множество правил, которое при этом будет охватывать малое количество объектов. Очевидно, что это приведёт к переобучению модели (явлению, при котором модель правильно принимает решения относительно данных из обучающей выборки, но плохо — относительно экземпляров, не участвовавших в обучении). Таким образом, чтобы избежать подобного явления и достичь хорошего результата классификации, необходимо установить, какие методы работы с несбалансированными данными наилучшим образом соответствуют текстовой специфике.

Согласно данным, которые приводят Хэ Хайбо и Эдуардо Гарсия за первое десятилетие двадцать первого века [4], число публикаций, посвящённых попытке решить проблему обучения модели на несбалансированных данных, постоянно растёт в значительном темпе. К примеру, с 2002 по 2007 год количество ежегодно публикуемых исследований выросло почти в 10 раз (с 10 до около 100). Учитывая, что

данным вопросом начали заниматься сравнительно недавно (тот же источник указывает, что первая работа появилась в 1997 году), а интерес к этой теме не спадает до сих пор, можно сделать вывод, что поиск решения этой проблемы имеет долгосрочные перспективы и будет оставаться актуальным и в будущем.

Результаты исследования. Чтобы проверить эмпирические методы для работы с несбалансированными данными, была выбрана задача бинарной классификации высказываний на русском языке в зависимости от того, корректны ли они или содержат какую-либо оскорбительную информацию (под ней понимается выраженное в текстовой форме оскорбление, то есть «унижение чести и достоинства человека в неприличной форме» [5]). Выбор задачи обусловлен её актуальностью (существующие решения данной проблемы в основной массе базируются на фильтрах определённых слов, а это не даёт возможности оценить контекст, в котором было использовано слово; кроме того, обычно списки стоп-слов для подобных фильтров включают очень ограниченное количество лексики) и принадлежностью к задаче несбалансированной классификации (очевидно, что оскорбительных высказываний значительно меньше, чем нейтральных текстов).

Для реализации модели, классифицирующей высказывания как оскорбительные или нейтральные, был выбран метод представления высказывания как вектора, компонентами которого будут являться слова-параметры, отобранные с помощью морфологического анализа и считающиеся значимыми для модели (так как данный список слов будет очищен от стоп-слов и служебных частей речи и лемматизирован). Значение компонентов вектора будет зависеть от того, встречается ли слово, ассоциированное с ним, в анализируемом высказывании.

После анализа методов сэмплинга, методов с распределением весов и методов на базе ядра и активном обучении, был сделан вывод о том, что

наиболее логично будет использовать метод с распределением весов или метод на базе ядра. Метод сэмплинга (метод, подразумевающий под собой добавление в обучающую выборку экземпляров менее представленного класса или удаление из неё экземпляров доминирующего класса с целью достичь сбалансированного распределения в наборе данных) был признан неподходящим для данной задачи, так как при его использовании слишком велик риск переобучения или потери информации. Данный подход во многом нацелен на то, чтобы изменить данные, а не сам классификатор непосредственно. Было доказано, что эффективнее работать с изменением порога решений и весами ошибок разного рода, чем пытаться изменить распределение обучающего набора данных случайным образом или синтетически [6].

Что касается оставшихся двух алгоритмов, то они имеют разные достоинства и недостатки: методы с распределением весов дают преимущество перед сэмплингом, работают с весами, штрафуя классификатор за ошибки при классификации объектов миноритарного класса. Это позволяет исказить классификатор в пользу менее представленного концепта. Помимо прочего, подходы с распределением весов имеют связь в теории с обучением на несбалансированных данных. Однако проблема вышеупомянутых техник в том, что они очень зависимы от матрицы весов, которую зачастую достаточно сложно вычислить. В то же время методы на базе ядра и активном обучении не имеют подобных проблем, но могут быть чрезмерно требовательны к сложности сопутствующих вычислений и могут совершенно игнорировать миноритарный класс из-за лежащего в их основе алгоритма машинного обучения SVM, который требует комбинации с методами сэмплинга и задания весов для приемлемого качества работы программы.

Для того чтобы обучить модель правильно классифицировать сообщения на предмет наличия в них оскорблений, были выбраны два

набора данных для обучающей выборки. В качестве примеров для класса оскорблений были взяты комментарии с таких новостных порталов, как newsland.ru, kp.ru, eho.msk.ru, svpressa.ru, и многих других за 2012–2013 годы, отмеченные экспертами как оскорбления по нескольким группам: «призыв к насилию и жестокости», «пропаганда культа насилия и жестокости» и «оправдание насилия и жестокости». Данные для класса нейтральных текстов были взяты из текстов синтаксически размеченного корпуса русского языка. Состав корпуса — научно-популярные, общественно-политические и информационные статьи из журналов и интернет-изданий (с 1980 года по настоящее время) и русская художественная проза второй половины XX в.

Вполне очевидно, что эта выборка данных неоднородна качественно: статьи/тексты и комментарии в интернете имеют особые отличительные черты. К примеру, как отмечает Е. В. Холодковская [7], «релевантными характеристиками интернет-комментария являются языковая краткость, меньшая нормированность, использование эмоционального синтаксиса, упрощение синтаксических структур и влияние синтаксиса разговорной речи». Из этого можно сделать предположение, что предложения из интернет-комментариев явно будут короче в среднем, чем те же структуры из текстов корпуса, и что это может породить неоднородность. А она, учитывая то, что предложения, как предполагается, будут собственно объектами классификации, может негативно повлиять на качество обучения модели. Для того чтобы её нивелировать, данные были подвергнуты процедуре фильтрования: все предложения, имеющие менее 40 символов, были удалены из выборки.

Таким образом, после данной процедуры средняя длина предложения в текстах интернет-комментариев и корпусных текстах составила примерно 111 и 125 символов соответственно, что является вполне допустимым показателем. Помимо прочего, данное фильтрование

помогло минимизировать влияние ошибок при разбиении текстов на предложения: сложные и неоднозначные для имеющегося инструментария случаи (например, такие как сокращения вида «и т. д.» и прямая речь автора), порождающие два предложения, были разрешены благодаря данной процедуре с минимальной потерей информации. В итоге после проведения фильтрации мощность классов стала равна 7039 и 1002 объектам для класса нейтральных высказываний и класса оскорблений соответственно, что означает пропорциональное соотношение между классами 7 к 1.

Для обучения модели были выбраны два наиболее часто используемых алгоритма машинного обучения для работы с несбалансированными данными: логистическая регрессия и метод опорных векторов с линейным ядром (Linear SVM). Реализации данных алгоритмов были взяты из библиотеки scikit-learn. Было обучено две модели для каждого алгоритма: невзвешенная и взвешенная (каждый из классов получил вес, равный единице, делённой на мощность класса). Они сравнивались на основе трёх метрик: точность, полнота и F-мера. Доля правильных решений классификатора (англ. accuracy) не использовалась для сопоставления моделей, так как при обучении на несбалансированных данных она не даёт никакой важной информации.

Модель, обученная с помощью алгоритма логистической регрессии: класс нейтральных высказываний

Класс нейтральных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная логистическая регрессия	0,98	**0,99**	**0,99**	700
Взвешенная логистическая регрессия	**0,99**	0,92	0,95	700

Таблица 1

Модель, обученная с помощью алгоритма логистической регрессии: класс оскорбительных высказываний

Класс оскорбительных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная логистическая регрессия	**0,95**	0,86	**0,90**	100
Взвешенная логистическая регрессия	0,74	**0,94**	0,83	100

Таблица 2

Модель, обученная с помощью алгоритма Linear SVM: класс нейтральных высказываний

Класс нейтральных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная Linear SVM	**0,99**	**0,95**	**0,97**	700
Взвешенная Linear SVM	**0,99**	0,94	0,96	700

Таблица 3

Модель, обученная с помощью алгоритма Linear SVM: класс оскорбительных высказываний

Класс оскорбительных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная Linear SVM	**0,73**	**0,94**	**0,82**	100
Взвешенная Linear SVM	0,70	**0,94**	0,80	100

Таблица 4

Выше приведены таблицы, показывающие качество работы моделей для умеренно несбалансированных данных. Несмотря на явную классовую несбалансированность данных (отношение между мощностями классов составляет 7 к 1), ни в одном из случаев не удалось добиться улучшения общего качества работы по сравнению с обучением на несбалансированных данных. Общий тренд таков, что применение весов для модели даёт прирост в полноте, однако снижает точность классификации. Для того чтобы проверить сделанный выше вывод, было решено обучить ещё одну пару моделей, используя те же алгоритмы, но

сделав выборку данных для класса оскорбительных высказываний в 2 раза меньше, чтобы несбалансированность мощностей классов проявилась ещё очевиднее (14 к 1 в пользу класса нейтральных высказываний).

Для обучения данных моделей было выбрано случайным образом 500 объектов из выборки класса оскорбительных высказываний. Тестовая выборка данных этого класса тоже была урезана в 2 раза и составила 50 экземпляров.

Модель, обученная с помощью алгоритма логистической регрессии на сильно несбалансированных данных: класс нейтральных высказываний

Класс нейтральных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная логистическая регрессия	0,96	**1,00**	**0,98**	700
Взвешенная логистическая регрессия	**0,98**	0,96	0,97	700

Таблица 5

Модель, обученная с помощью алгоритма логистической регрессии на сильно несбалансированных данных: класс оскорбительных высказываний

Класс оскорбительных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная логистическая регрессия	**1,00**	0,48	0,65	50
Взвешенная логистическая регрессия	0,68	**0,76**	**0,72**	50

Таблица 6

Модель, обученная с помощью алгоритма LinearSVM на сильно несбалансированных данных: класс нейтральных высказываний

Класс нейтральных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная Linear SVM	**0,98**	**0,98**	**0,98**	700
Взвешенная Linear SVM	**0,98**	0,96	0,97	700

Таблица 7

Модель, обученная с помощью алгоритма LinearSVM на сильно несбалансированных данных: класс оскорбительных высказываний

Класс оскорбительных высказываний	Точность	Полнота	F-мера	Количество объектов в тестовой выборке класса
Невзвешенная Linear SVM	**0,69**	0,68	0,69	50
Взвешенная Linear SVM	0,65	**0,76**	**0,70**	50

Таблица 8

После проведения подобного испытания на сильно несбалансированных данных (отношение между мощностями классов 14 к 1), было установлено, что падение показателя точности невзвешенных моделей действительно приводит к относительному преимуществу по результатам классификации (пусть и незначительному для LinearSVM моделей).

Что касается показателей полноты и точности в отдельности, то результаты тестирования моделей показали, что невзвешенные модели имеют более высокий показатель полноты при классификации нейтральных высказываний, в то время как использование взвешенных моделей для этого класса даёт преимущество в точности (однако следует заметить, что данное преимущество незначительно и не позволяет улучшить качество классификации в целом).

При классификации оскорбительных высказываний были получены противоположные предыдущему случаю результаты: использование взвешенных моделей помогает значительно повысить показатель полноты, а использование невзвешенных — точности.

Выводы. В результате анализа методов классификации для несбалансированных данных было установлено, что эффективнее работать

с изменением порога решений и весами ошибок разного рода, чем пытаться изменить распределение обучающего набора данных случайным образом или синтетически. В связи с этим наиболее эффективным методом для работы с несбалансированными данными был признан метод с распределением весов, так как он работает с весами и имеет связь в теории с обучением на несбалансированных данных.

В результате анализа результатов было выяснено, что использование методов с распределением весов не помогло улучшить результат классификации текстов по сравнению с обучением на умеренно несбалансированных данных для данной задачи. Однако после проведения подобного испытания на сильно несбалансированных данных, было установлено, что падение показателя точности невзвешенных моделей действительно приводит к относительному преимуществу по результатам классификации (пусть и незначительному для LinearSVM моделей). Также было выявлено, что во всех проведённых испытаниях невзвешенные модели имели более высокий показатель полноты при классификации нейтральных высказываний, в то время как использование взвешенных моделей для этого класса даёт преимущество в точности. При классификации же оскорбительных высказываний были получены противоположные предыдущему случаю результаты: использование взвешенных моделей помогает значительно повысить показатель полноты, а использование невзвешенных — точности.

Список литературы

1. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. The MIT Press, 2012. 432 pp.
2. Вентцель Е. С. Теория вероятностей: Учеб. для вузов. 6-е изд. стер. М.: Высшая школа, 1999. 576 с.
3. Provost F. Machine Learning from Imbalanced Data Sets: Proceedings of the AAAI’2000 workshop on imbalanced data sets // AAAI URL: http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf (дата обращения: 16.11.2016).

4. He H., Garcia E. A. Learning from Imbalanced Data // IEEE transactions on knowledge and data engineering, vol. 21, no. 9, September 2009. Pp. 1263–1284.

5. Лесниевски-Костарева Т. А. Словарь-справочник уголовного права. М.: НОРМА-ИНФРА-М, 2000. 422 с.

6. Liu X., Zhou Z. The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study // Proc. Int'l Conf. Data Mining, 2006. Pp. 970–974.

7. Холодковская Е. В. Особенности синтаксиса англоязычного интернет-коммментария социальной сети Facebook // Вестник Волгоградского государственного университета. Серия 2: Языкознание. 2014. №1. С. 79–83. URL: http://cyberleninka.ru/article/n/osobennosti-sintaksisa-angloyazychnogo-internet-kommentariya-sotsialnoy-seti-facebook (дата обращения: 16.11.2016).

References

1. Mohri M., Rostamizadeh A., & Talwalkar A. (2012). Foundations of Machine Learning. The MIT Press.

2. Venttsel Ye. S. (1999) Teoriya veroyatnostey [Probability theory]. Moscow: Vysshaya shkola.

3. Provost F. (2000). Machine Learning from Imbalanced Data Sets, Proceedings of the AAAI'2000 workshop on imbalanced data sets (http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf).

4. He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284.

5. Lesniyevski-Kostareva T. A. (2000) Slovar-spravochnik ugolovnogo prava [Glossary of Criminal Law]. Moscow: Norma-Infra-M.

6. Liu, X., & Zhou, Z. (2006). The Influence of Class Imbalance on Cost-Sensitive Learning: An Empirical Study. *Proc. Int'l Conf. Data Mining*, 970–974.

7. Holodkovskaya, E. V. (2014). Syntactic Peculiarities of English Internet Comments of Facebook Social Network. *Science Journal of Volgograd State University. Linguistics*, *2014*(1), 79–83. Retrieved from http://cyberleninka.ru/article/n/osobennosti-sintaksisa-angloyazychnogo-internet-kommentariya-sotsialnoy-seti-facebook