А.Д. ВАРЛАМОВ

Проблемы практического применения наивного байесовского классификатора для оценки тональности текстов

УДК 004.912

Муромский институт (филиал) ФГБОУ ВО «Владимирский государственный университет имени А.Г. и Н.Г. Столетовых», г. Муром

В статье описаны проблемы применения на практике байесовского классификатора для оценки тональности текстов. Предложены пути решения данных проблем, которые реализованы в двух разработанных алгоритмах. Проведены исследования этих алгоритмов совместно с реализованным классическим байесовским классификатором.

Введение

В настоящее время в интернете можно встретить огромное количество отзывов, комментариев и мнений о товарах и услугах. Автоматическая обработка этих данных позволит компаниям, интернет магазинам и различным сетевым сервисам оценить удовлетворенность клиентов приобретением товаров и использованием услуг. При наличии большого количества оставленных постов пользователи смогут быстро получить общую объективную оценку по количественной шкале об интересующем его объекте, а интернет магазины получат возможность ее использовать в реализации рекомендательных функций. Удовлетворенность потребителей отражается тональностью их высказываний в отзывах.

Под тональностью текста понимается эмоциональный окрас и эмоциональная оценка автора по отношению к объектам, речь о которых идёт в тексте. Авторские высказывания могут иметь позитивный или негативный тон. Поэтому можно выделить классы документов с положительной и с отрицательной тональностью

текста, а также класс, в котором нет явного превосходства одного из них.

Анализ тональности (sentiment analysis) относится к типу задач классификации объектов [2,4-9]. Чаще всего текстовые документы делят на два класса (позитивный и негативный тон) и на три класса (с учетом нейтрально-тональных текстов). Одним из наиболее известных методов классификации текстов на основе машинного обучения является наивный байесовский классификатор (Naive Bayes classifier) [10]. Исследование публикаций показало, что подавляющее большинство современных работ, посвященных анализу мнений в текстах [1,2,6-9,11,12], в анализе подходов к реализации и (или) сравнении результатов, не обходят стороной традиционный байесовский метод. Отметим, что он представляет собой классическую, теоретически обусловленную вероятностную модель, реализовать которую не составляет большого труда.

Но, как известно, теоретически обусловленные модели не всегда пригодны для решения практических задач или их реализация сопровождается большими трудностями. Далее мы теоретическими выкладками обоснуем критику классического байесовского подхода для эмоциональных оценок текстов.

Наивный байесовский классификатор текстов на группы тональности

Возьмем самый простой классификатор – разделение документов на два класса. Байесовский метод предполагает сбор статистики на основе заранее подготовленных текстов и ее дальнейшее использование. Итак, анализом обучающей коллекции текстов можно рассчитать следующие величины:

$P(T)$ - вероятность того, что встретится документ с положительной характеристикой;

$P(N)$ - вероятность того, что встретится документ, сопровождающийся преимущественно отрицательными эмоциями;

$P(S|T)$ - вероятность того, что в положительном документе встретится слово S;

$P(S|N)$ - вероятность того, что в отрицательном документе встретится слово S.

На основе наивного байесовского классификатора тональность текста оценивается по формуле:

$$\hat{P}(T) = \frac{P(T)\prod_{i=1}^{I} P(S_i \mid T)}{P(T)\prod_{i=1}^{I} P(S_i \mid T) + P(N)\prod_{i=1}^{I} P(S_i \mid N)} \quad (1)$$

где I - количество слов в документе, S_i - i-е слово в документе. При значении $\hat{P}(T) > 0{,}5$, тональность текста оценивается как положительная, иначе как отрицательная.

Как мы видим, оценить тональность текста по данной формуле достаточно просто – все величины, входящие в нее, рассчитываются на этапе анализа обучающей выборки. При практическом внедрении обученного классификатора величины $P(T)$ и $P(N)$ становятся константами, а $P(S_i|T)$ и $P(S_i|N)$ могут быть сохранены в таблицу базы данных. Оценка тональности проводится следующим образом: документ разбивается на слова, для каждого из которых из базы берутся условные вероятности, и рассчитывается тональность по формуле (1).

Но, не смотря на теоретическую обусловленность этой модели, на практике возникает ряд проблем, которые описаны ниже.

Обоснование проблем применения байесовского классификатора

Проблема 1 связана с погрешностями статистических оценок и вычислений.

Ошибки статистических оценок появляются при расчете значений $P(S_i|T)$ и $P(S_i|N)$, так как статистические величины всегда приближены (не равны) вероятностным. Но их можно уменьшить, увеличивая объем обучающей выборки. Средняя абсолютная ошибка определения вероятности некоторой случайной величины по частоте ее встречаемости пропорциональна среднеквадратическому отклонению, которое, в свою очередь, обратно пропорционально квадратному корню из объема выборки. Поэтому ошибки статистических оценок и объем выборки соотносятся следующим образом:

$$E_S \sim \frac{1}{\sqrt{D}} \quad (2)$$

где E_s - средняя абсолютная ошибка, относящаяся к величинам $P(S_i|T)$ и $P(S_i|N)$; D - количество документов в обучающей выборке.

Погрешность вычисления связана с использованием формулы (1). Можно показать, что ошибка определения тональности байесовским классификатором пропорциональна количеству слов оцениваемого текста. Из теории погрешностей известно, что о*тносительная погрешность произведения* равна сумме относительных погрешностей отдельных сомножителей [3]. Количество сомножителей в формуле (1) пропорционально количеству слов документа, а значит

$$E_П \sim I_{ср}. \quad (3)$$

где $E_П$ - средняя относительная ошибка произведений в формуле (1), $I_{ср}$ - среднее количество слов, входящих в текстовый документ.

Если учесть все арифметические операции в (1), можно показать, что

$$E \sim E_П, \quad (4)$$

где E - средняя абсолютная ошибка оценки тональности текста. Очевидно, также, что

$$E \sim E_s. \quad (5)$$

Из (2), (3), (4) и (5) получаем

$$E \sim \frac{I_{ср}}{\sqrt{D}}. \quad (6)$$

Вычислительная сложность W алгоритма обработки обучающей выборки, а также трудоемкость работы экспертов, создающих эту выборку, пропорциональны количеству слов выборки:

$$W \sim D\, I_{ср}. \quad (7)$$

На практике требуется обеспечить заданную минимальную погрешность E итоговой оценки $\hat{P}(T)$. Тогда величину ошибки E следует рассматривать как постоянную и из формул (6) и (7) получим соотношение

$$W \sim I_{ср}^3, \text{ при } E=\text{const}. \quad (8)$$

Таким образом, можно заключить, что обеспечение требуемого уровня точности оценки (E=const) выполняется за время, пропорциональное третьей степени среднего количества слов в анализируемых текстах. С точки зрения теории алгоритмов вычислительная задача, имеющая полиномиальный порядок

временной сложности третьей степени, является неэффективной и не желательной в реализации, если используются данные большого размера. Применительно к описываемой задаче можно заключить, что использование наивного байесовского классификатора для оценки тональности текстовых документов с большим количеством слов не целесообразно.

Также следует подчеркнуть, что большой словарный запас русского языка и без того требует выборки, состоящей из десятков тысяч документов, а разная частота встречаемости слов в текстах увеличивает погрешности статистических оценок многих слов.

Проблема 2. Учет редких слов в модели сильно искажает оценку тональности.

Чем меньше частота встречаемости слова в текстах, чем реже оно будет попадать в положительный и отрицательный классы при статистической обработке обучающей коллекции. Частоты встречаемости такого слова в документах разной тональности менее точно будут соответствовать вероятностям $P(S_i|T)$ и $P(S_i|N)$. Полученная статистическая ошибка является погрешностью E_s для одного слова и может значительно увеличить значение E при агрегации.

Если учитывать очень редкие слова, а также слова, в которых допущена ошибка или опечатка, могут быть получены совсем неадекватные результаты. Например, если в обучающей выборке некоторое слово встретилось только в положительных текстах, а другое только в отрицательных (это вполне возможно для слов с очень низкой частотой встречаемости), и эти два слова встречаются в новом оцениваемом тексте, то в формуле (1) возникает неопределенность типа 0/0. А если только одно из этих слов встретится в документе, то оно и определит оценку тональности не зависимо от всех других слов документа. Поэтому в наивном байесовском классификаторе следует использовать ограниченный словарь, исключая из него редкие слова.

Проблема 3. Не удобная программная реализация.

Эта проблема относится только к аспекту программирования. При практической реализации значения $P(S_i|T)$ и $P(S_i|N)$ будут храниться в базе данных (возможно и в других источниках структурированных данных). Среди стандартных агрегирующих

функций в SQL-запросах к базам данных нет произведения (как в прочем и в других инструментах работы со структурированными данными). Поэтому единым запросом к базе не может быть получено произведение выбранных значений ни в исследовательских проектах, ни в рабочих вариантах системы. Это ведет к усложнению логики программного кода и времени работы алгоритмов.

Решение проблем.

Чтобы избежать указанных проблем, можно разбить текст на фрагменты (например, на абзацы, на предложения или на слова) и определить тональность каждого фрагмента. Затем среднеарифметическим полученных значений или их взвешенной суммой оценить тональность всего документа. Если использовать самую высокую степень разбиения (на слова), формула оценки тональности будет иметь вид:

$$\hat{P}(T) = \frac{1}{I}\sum_{i=1}^{I}\hat{P}(T \mid S_i), \tag{9}$$

где $\hat{P}(T \mid S_i)$ - лексическая тональность, выраженная на уровне слова S_i.

К данной формуле применимо уже другое правило определения величины ошибки: если слагаемые одного и того же знака, то относительная погрешность их суммы не превышает наибольшей из предельных относительных погрешностей слагаемых [3]. Учитывая данный факт, сложность работ для обеспечения требуемого уровня точности оценки тональности оказывается на порядок меньше по сравнению с Байесовским подходом.

Данный подход снимает также вторую (балансируется влияние слов на оценку тональности) и третью (сумма - один из самых распространенных агрегатов) проблемы, обозначенные в статье.

Результаты исследований

Чтобы подтвердить неточности классического байесовского подхода для оценки тональности больших текстов, было проведено исследование, для которого реализованы три алгоритма: классический наивный байесовский классификатор, оценивающий тональность по формуле (1); усреднение байесовских оценок,

рассчитанных для предложений; усреднение влияния на тональность каждого слова, реализованное на основе формулы (9).

В качестве набора исходных данных использовались коллекции семинара РОМИП [9] с отзывами о книгах, фильмах и фотокамерах, разделенные экспертами на две группы: положительные и отрицательные. Вся совокупность текстов разделена на обучающее и тестовое множество. Первое использовалось для машинного обучения, то есть для определения статистических величин, необходимых для реализации алгоритмов (лексических тональностей слов, характеристик выборки и так далее). Второе затем применялось для оценки качества работы алгоритмов.

Количество документов в обучающей (для расчета статистики) выборке составило 21692 штуки. Тестовая выборка содержит такой же объем текстов. Так как требовалось исследовать алгоритмы на данных разного объема, она была разбита на две выборки. В первую вошли самые короткие тексты (с количеством символов до 999), а во вторую собраны длинные отзывы (не менее 1000 символов на документ) . Средний размер документов обучающей выборки составляет 33 слова, около 5 предложений. Это характерно для многих аннотаций документов, блогов, новостей. Характеристики всех коллекций, задействованных в исследованиях, приведены в таблице 1.

Таблица 1

Характеристики текстовых коллекций

Выборка	Назначение	Количество текстов	Среднее количество слов в тексте
Разные тексты	обучающая	21692	33
Короткие тексты	тестовая	19169	23
Длинные тексты	тестовая	2523	113

Теперь, сопоставляя оценки экспертов и результаты работы алгоритмов, можно получить объективные показатели их качества. Для каждого алгоритма рассчитан набор метрик, рекомендованных семинаром РОМИП [9] (таблица 2). Значения метрик соответствуют задаче поиска положительных отзывов обученными алгоритмами.

Таблица 2

Сравнение методов при оценке тональности

Метрика	Классификатор Байеса		Усреднение байесовских оценок для предложений.		Усреднение лексических тональностей	
Выборка	Короткие тексты	Длинные тексты	Короткие тексты	Длинные тексты	Короткие тексты	Длинные тексты
Полнота	97,25	97,66	97,67	99,95	96,85	99,65
Точность	99,42	97,38	97,80	99,98	99,66	99,96
Аккуратность	96,96	95,16	95,85	99,97	96,81	99,64
F-мера	98,33	97,39	97,73	99,96	98,24	99,80
Процент документов с не определенной тональностью.	1,23	53,01	22,60	0,12	0	0

Полученные результаты подтвердили, что пригодность классификатора Байеса для длинных текстов становится очень низкой ввиду описанных в статье проблем. Для больших текстов целесообразно применять данный метод не целиком ко всему документу, а отдельно к его фрагментам (например, к предложениям) с дальнейшим сведением результатов к общей оценке. Для коротких отзывов результаты исследования свидетельствуют об обратном: усреднение байесовских оценок дало худшую характеристику (ввиду того, что малое четное количество предложений часто давало равновероятную оценку принадлежности записи к обеим группам).

Кроме метрик в число показателей мы ввели количество документов, тональность которых не удалось определить. Это тексты, при оценке которых возникла неопределенность типа 0/0 или оцененная вероятность точно равнялась 0,5. Лучшее значение этого критерия показал метод усреднения лексических тональностей. Классификатор Байеса - единственный из рассмотренных методов, который не исключает появления неопределенности 0/0 (вероятность которой возрастает с повышением количества слов в тексте). Усреднение байесовских оценок дает равновероятный результат классификации, если количество предложений с противоположными результатами оценки тона одинаково.

Таким образом, мы рекомендуем применять предложенный подход для оценки тональности текстов большого объема. Стоит также иметь ввиду, что описанные проблемы и подход к их решению актуальны не только для байесовской оценки эмоций, но и в целом для задач классификации текстов.

Литература

1. Брунова. Е.Г. Методика составления оценочного лексикона для контент-анализа мнений. 2012

2. Васильев В.Г., Худякова М.В., Давыдов С. Классификация отзывов пользователей с использованием фрагментных правил // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая – 3 июня 2012 г.). Вып. 11: В 2 т. Т. 2- М: Изд-во РГГУ, 2012. С. 66-76.

3. Выгодский М.Я. Справочник по элементарной математике. М.: АСТ Астрель, 2006. - 509с.

4. Евстигнеева О.И., Садыков С.С., Суслова Е.Е., Белякова А.С. Критерии выделения групп риска из лиц трудоспособного возраста при медицинских исследованиях на системе АСПО // Алгоритмы, методы и системы обработки данных. 2012. № 19. С. 33-39.

5. Еремеев С.В., Ковалев Ю.А. Алгоритм классификации пространственных объектов на основе модели Random forest // Алгоритмы, методы и системы обработки данных. 2017. № 35. С. 9-15.

6. Котельников Е.В., Клековкина М.В. Автоматический анализ тональности текстов на основе методов машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по матер. ежегодн. Междунар. конф. «Диалог». 2012. № 11 (18). С. 753–762.

7. Ленкин А.В., Баженов Р.И. Исследование систем для Text Mining // Постулат. 2017. № 1 (15). С. 3.

8. Юсупова Н.И., Богданова Д.Р., Бойко М.В. Алгоритмическое и программное обеспечение для анализа тональности текстовых сообщений с использованием машинного обучения // Вестник Уфимского государственного авиационного технического университета, 2012, С. 91-99.

9. Российский семинар по Оценке Методов Информационного Поиска (РОМИП) [Электронный ресурс]. — Режим доступа: http://www.romip.ru.

10. Lewis D.D. Naive (Bayes) at forty: The independence assumption in information retrieval. Proceedings of 10th European Conference on Machine Learning, Chemnitz, Germany, 1998, pp. 4–15.

11. Pavlova O.V., Bazhenov R.I., Bogachenko N.G., Salnikova Y.A., Guryan N.V. Comprehension of Chinese idiomatic expressions in word-for-word translation into the Russian language / Information. 2016. Т. 19. № 6A. С. 1845-1852.

12. Poroshin V. Proof of concept statistical sentiment classification at ROMIP 2011 // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог». Вып. 11 (18), М.: Изд-во РГГУ, 2012, С. 60–65.