Последовательный алгоритм для обучения классификаторов текстов

Дэвид Д. Льюис (lewis@research.att.com) и Уильям А. Гейл (gale@research.att.com)

AT&T Bell Laboratories; Мюррей-Хилл, Нью-Джерси 07974; США

В книге В. Брюса Крофта и К. Дж. ван Райсбергена, ред. SIGIR 94: Труды семнадцатой ежегодной международной конференции ACM-SIGIR по исследованиям и разработкам в области информационного поиска, Springer-Verlag, London, pp. 3{12.

Аннотация

Возможность дешевого обучения текстовых классификаторов имеет решающее значение для их использования в информационном поиске, контент-анализе, обработке естественного языка и других задачах с данными, частично или полностью состоящими из текста. Алгоритм последовательной выборки при машинном обучении статистических классификаторов был разработан и протестирован на задаче категоризации текстов новостных лент. Этот метод, который мы назвали выборкой неопределенности, позволил сократить в 500 раз объем обучающих данных, которые необходимо было бы классифицировать вручную для достижения заданного уровня эффективности.

1 Введение

Классификация текстов - это автоматизированная группировка текстовых или частично текстовых объектов. Поиск документов, категоризация, маршрутизация, сортировка и кластеризация, а также задачи обработки естественного языка, такие как тегирование, разбор слов по смыслу и некоторые аспекты понимания, могут быть сформулированы как классификация текста. По мере увеличения объема онлайновых текстов растет потребность в классификации текстов для помощи в их анализе и управлении ими.

Одно из преимуществ формулировки задач обработки текста как классификации заключается в том, что методы статистики и машинного обучения могут быть использованы для автоматического формирования классификаторов текста. Хотя использование машинного обучения требует ручного аннотирования обучающих данных метками классов, такое аннотирование требует меньше навыков и затрат, чем, например, построение правил классификации вручную [1].

Зачастую объем доступного текста больше, чем экономически выгодно маркировать, поэтому для маркировки необходимо выбрать подмножество или выборку данных.[1] Случайная выборка [3] обычно не является эффективной. Если только 1 из 1000 Тексты являются членами класса (не нетипичными), и только 500 текстов могут быть помечены, то случайная выборка обычно будет содержать 500 отрицательных примеров и ни одного положительного. Это не поможет обучить классификатор отличать положительные примеры от отрицательных.

Обратная связь по релевантности [4] представляет собой разновидность неслучайной выборки. В этом случае пользователей просят пометить те тексты, которые, по мнению текущего классификатора, с наибольшей вероятностью являются членами класса. Такой подход, который мы можем назвать выборкой релевантности, является разумной стратегией в контексте текстового поиска, где пользователь больше заинтересован в просмотре релевантных текстов, чем в эффективности итогового классификатора. Обратная связь по релевантности также была предложена для поиска примеров необычных смыслов слов [5]. Однако обратная связь по релевантности имеет много проблем как подход к выборке. Она работает все хуже по мере совершенствования классификатора и подвержена отбору избыточных примеров.

Выборка релевантности - это последовательный подход к выборке, поскольку маркировка более ранних примеров приводит к выбору более поздних [6]. В данной статье описывается альтернативный последовательный подход, выборка неопределенности, мотивированный результатами теории вычислительного обучения. Выборка неопределенности - это итерационный процесс ручной маркировки примеров, классификации на основе этих примеров и использования классификатора для отбора новых примеров, принадлежность которых к определенному классу неясна. Мы показываем, как можно обучить вероятностный классификатор, используя выборку неопределенностей. Проверка этого метода на задаче категоризации текста показала сокращение до 500 раз количества примеров, которые должны быть помечены для получения классификатора с заданной эффективностью.

2 Обучение с помощью запросов

Классификатор часто можно выучить на меньшем количестве примеров, если позволить алгоритму обучения создавать искусственные примеры или запросы на членство и просить учителя обозначить их [7, 8].[2] Во многих задачах обучения создание искусственных примеров не представляет проблемы. Однако художественный текст, созданный алгоритмом обучения, вряд ли будет легитимным выражением на естественном языке и, вероятно, не будет интерпретирован человеком-преподавателем.

---

[1] Исключение составляют случаи, когда доступны большие объемы ранее помеченного текста, например, когда автоматизированная категоризация текста используется для замены или помощи существующей группе ручных индексаторов [2].

[2] В этой статье под "\queries" всегда подразумеваются запросы на членство, а не запросы на поиск текста.

1. Создайте начальную классификацию
2. В то время как учитель готов обозначить примеры
   (a) Примените текущую классификацию к каждому немаркированному примеру
   (b) Найдите b примеров, для которых классификатор наименее уверен в принадлежности к классу
   (c) Попросите учителя пометить подвыборку примеров b.
   (d) Обучение новой классификации на всех помеченных примерах

Рисунок 1. Алгоритм выборки неопределенности с одним классификатором.

Недавно было предложено несколько алгоритмов обучения с помощью запросов, которые используют существующие примеры, а не создают искусственные [9, 10, 11]. Эти алгоритмы просят учителя помечать только те примеры, принадлежность которых к тому или иному классу достаточно \неопределенна". Было использовано несколько определений неопределенности, но все они основаны на оценке того, насколько вероятно, что классификатор, обученный на ранее помеченных данных, выдаст правильную метку класса для данного немеченого примера. Рассматривая этот метод как метод выборки, а не запроса, мы называем этот подход выборкой неопределенности.

Сеунг, Оппер и Сомполински [11] представили теоретический анализ \query by committee" (QBC), алгоритм, который для каждого немаркированного примера случайным образом выбирает два классификатора из пространства версий, т.е. множество всех классификаторов, согласующихся с помеченными обучающими данными [12]. Предполагается постоянный поток немеченых данных, из которого QBC запрашивает у учителя метки классов только для тех примеров, для которых два выбранных классификатора расходятся.

Фройнд, Сеунг, Шамир и Тишби распространили результат QBC на широкий спектр классификационных форм [13]. Они доказывают, что при определенных предположениях количество запросов, сделанных после изучения m случайных примеров, будет логарифмически возрастать от m, а ошибка обобщения будет уменьшаться почти так же быстро, как если бы запросы делались на всех примерах. Точнее, ошибка обобщения уменьшается как O(1=m). Таким образом, с точки зрения количества запросов ошибка обобщения уменьшается экспоненциально быстро.

Это провокационный результат, поскольку он подразумевает, что эффект от обучения на помеченных данных может быть получен за счет затрат на получение немеченых данных и маркировку лишь логарифмической их части. Однако предположения QBC включают в себя отсутствие шума в данных, существование идеального детерминированного классификатора и возможность случайного выбора классификаторов из пространства версий, что является проблематичным для задач реального мира. Эффективность QBC и других методов для задач реального мира еще предстоит определить.

Эвристическая альтернатива случайному выбору классификаторов из пространства версий в QBC заключается в том, чтобы позволить алгоритму обучения делать то, что он всегда делает, - выбирать одного классификатора из пространства версий. Если классификатор может не только принимать классификационные решения, но и оценивать их определенность, оценка определенности может быть использована для выбора примеров.

Подход с использованием одного классификатора к выборке неопределенности имеет ряд теоретических недостатков, включая заниженную оценку истинной неопределенности и погрешности, вызванные нерепрезентативными классификаторами [9, 10]. С другой стороны, эксперименты с использованием одного классификатора для составления произвольных запросов [14] или выбора подмножеств помеченных данных [8, 15] показали значительное ускорение обучения. Выборка релевантности, которая оказалась весьма эффективной для текстового поиска, также использует один классификатор.

3 Алгоритм выборки из неопределенности

На рисунке 3 представлен алгоритм выборки неопределенности из неограниченного набора примеров с помощью одного классификатора. В идеале b, количество примеров, отбираемых на каждой итерации, должно быть равно 1, но большие значения могут быть уместны, если оценка и отбор примеров требуют больших затрат. Этот алгоритм можно использовать с любым типом классификатора, который одновременно предсказывает класс и дает оценку того, насколько уверенным является это предсказание. Вероятностные, нечеткие, ближайшие соседи и нейронные классификаторы, а также многие другие, удовлетворяют этому критерию или могут быть легко модифицированы для этого. Возможно, самым важным требованием является то, что измерения относительной уверенности должны производиться даже тогда, когда классификатор был сформирован на очень небольшом количестве обучающих примеров.

Выборка неопределенности похожа на стратегию обучения на неправильно классифицированных экземплярах [16, 17]. Разница заключается в том, что, когда данные не помечены, мы должны использовать сам классификатор, чтобы догадаться, какие примеры неправильно классифицированы. Отметим, что начальный классификатор играет важную роль, поскольку без него может пройти длительный период случайной выборки, прежде чем будут найдены примеры низкочастотного класса.

4 Вероятностная классификация текстов

В этом разделе мы описываем классификационную форму, которая производит оценки $P(C_i \mid w)$, апостериорной вероятности того, что пример с паттерном w принадлежит к классу C_i . Оценки этой вероятности могут использоваться как для принятия решения о том, когда пример должен быть отнесен к классу, так и для оценки того, насколько вероятно, что классификация

будет правильным. Мы описываем, как обучается классификатор и как мы используем его для выборки неопределенностей и классификации.

4.1 Вероятностная классификация

Классификаторы, которые оценивают апостериорную вероятность по правилу Байеса:

$$P(C_i|w) = \frac{P(w|C_i)\,P(C)_i}{\sum_{j=1} P(w|C_j)\,P(C)_j} \quad (1)$$

были применены к различным задачам классификации текстов, включая текстовый поиск [18], категоризацию текстов [19, 20], и идентификация смысла слов [5]. Здесь C_i - это расходящийся и исчерпывающий набор классов, к которым может принадлежать пример, а $w = (w_1 ; :::; w_d)$ - наблюдаемый паттерн.[3] $P(w|C_i)$ - это условная вероятность того, что пример имеет паттерн w, учитывая, что он принадлежит классу C_i , а $P(C_i)$ - это предварительная вероятность того, что пример принадлежит классу C_i.

В данной работе мы рассматриваем только случай q = 2, поэтому существует два класса $C_1 = C$ и $C_2 = \bar{C}$, причем $P(\bar{C}) = 1 - P(C)$. В этом случае полезно выразить относительные апостериорные вероятности C и $\bar{C}$ в виде отношения шансов:

$$\frac{P(C|w)}{P(\bar{C}|w)} = \frac{P(C)}{P(\bar{C})}\,\frac{P(w|C)}{P(w|\bar{C})} \quad (2)$$

Учитывая огромное количество возможных w, оценивать $P(w|C)$ путем прямого наблюдения за w в обучающем множестве бесполезно. Приняв определенные предположения о независимости [21], мы можем сделать следующую декомпозицию:

$$\frac{P(C|w)}{P(\bar{C}|w)} = \frac{P(C)}{P(\bar{C})}\prod_{i=1}^{d}\frac{P(w_i|C)}{P(w_i|\bar{C})} \quad (3)$$

Затем, используя тот факт, что $P(\bar{C}|w) = 1 - P(C|w)$, плюс некоторые арифметические манипуляции, мы можем получить следующее выражение для $P(C|w)$:

$$P(C|w) = \frac{\exp\left(\log\frac{P(C)}{1-P(C)} + \sum_{i=1}^{d}\log\frac{P(w_i|C)}{P(w_i|\bar{C})}\right)}{1 + \exp\left(\log\frac{P(C)}{1-P(C)} + \sum_{i=1}^{d}\log\frac{P(w_i|C)}{P(w_i|\bar{C})}\right)} \quad (4)$$

Уравнение 4 редко используется непосредственно при классификации текстов, вероятно, потому, что его оценки $P(C|w)$ систематически неточны. Одна из причин такой неточности заключается в том, что предположения о независимости, сделанные при составлении уравнения 3, всегда неверны, когда w_i 's являются словами или другими признаками, полученными из естественного языка. Другая проблема заключается в том, что P (C) обычно мала и, следовательно, ее трудно оценить, и эта проблема усугубляется, когда обучающее множество не является случайной выборкой.

Логистическая регрессия [22] обеспечивает частичное решение этих проблем. Это общая техника объединения нескольких значений предикторов для оценки апостериорной вероятности. Оценка имеет следующий вид:

$$P(C|x) = \frac{\exp(a + b\,x_{11} + ::: + b\,x_{mm})}{1 + \exp(a + b\,x_{11} + ::: + b\,x_{mm})} \quad (5)$$

Было предложено несколько подходов к использованию логистической регрессии для классификации текстов [23, 24, 25]. Сходство между уравнением 4 и уравнением 5 побудило нас попробовать особенно простой подход, в котором в качестве единственной предикторной переменной используется логарифмическое отношение правдоподобия из формулы независимости Байеса:

$$P(C|w) = \frac{\exp\left(a + b\sum_{i=1}^{d}\log\frac{P(w_i|C)}{P(w_i|\bar{C})}\right)}{1 + \exp\left(a + b\sum_{i=1}^{d}\log\frac{P(w_i|C)}{P(w_i|\bar{C})}\right)} \quad (6)$$

Интуитивно мы могли бы надеяться, что логистический параметр a заменит труднооцениваемые предварительные логарифмические шансы в уравнении 4, а b будет служить для смягчения экстремальных логарифмических коэффициентов правдоподобия, возникающих из-за нарушения независимости. Мы действительно обнаружили, что эта простая формулировка хорошо работает для категоризации текста, хотя мы не сравнивали ее с более сложными формулировками, предложенными другими авторами. Заметим, что наш подход, вероятно, не подойдет, если документы будут сильно различаться по длине.

---

[3] *Мы стараемся отличать пример e от соответствующего образца w, поскольку разные примеры могут иметь одинаковые значения признаков $w_1 ; ::: ; w_d$*

4.2 Обучение классификатора

Первым шагом в использовании уравнения 6 является оценка значений P (w_i jC)=P (w_i jC Мы использовали следующую оценку:

$$\frac{P(w_i jC) :}{P_i(w jC} = \frac{-\frac{cpi+(Np+0:5)=(Np+Nn+1)}{Np+d(Np+0:5)=(Np+Nn+1)}}{-\frac{cni+(Nn+0:5)=(Np+Nn+1)\ Nn}{+d(Nn+0:5)=(Np+Nn+1)}} \qquad (7)$$

Здесь N_p и N_n - количество лексем в положительном и отрицательном обучающих наборах соответственно, c_{pi} и c_{ni} - количество экземпляров w_i в положительном и отрицательном обучающих наборах соответственно, а d - количество признаков. Это специальная оценка, слабо обоснованная аналогией с оценкой ожидаемого правдоподобия [26]. Приведенная выше оценка позволяет избежать экстремальных оценок отношения правдоподобия, когда N_p и N_n имеют очень разные размеры, например, до того, как наша процедура выборки начнет н а х о д и т ь положительные примеры.

При классификации текстов обычно имеется огромное множество потенциальных w_i 's, например, все типы (отдельные слова) в коллекции документов. Использование отбора признаков для уменьшения этого набора (или, что эквивалентно, для фиксации всех значений, кроме нескольких, на уровне 0) может повысить эффективность [19]. В качестве меры качества признаков мы использовали:

$$(c_{pi} + c_{ni}) \log : \frac{P(w_{ij}C)}{P(w_i jC} \qquad (8)$$

Мы отбирали признаки в порядке возрастания этого значения до тех пор, пока не была достигнута определенная доля (0,7 в представленных здесь экспериментах) от общего числа баллов всех обучающих примеров. Это делалось отдельно для признаков с положительными и отрицательными коэффициентами правдоподобия.

После отбора признаков значения логарифмического правдоподобия используются для вычисления:

$$\sum_{i=1}^{d} \log \frac{P(w_i jC)}{P(w_i jC} \qquad (9)$$

для каждого обучающего примера. Логистическая регрессия используется для нахождения значений a и b, которые дают наилучший результат.

t этого значения к вероятности принадлежности к классу.

4.3 Выборка неопределенности с помощью вероятностной классификации

Выборка неопределенности проста при наличии классификатора, который оценивает P (Cjw). На каждой итерации к каждому примеру можно применить текущую версию классификатора и отобрать примеры с оценкой P (Cjw), близкой к 0:5, поскольку 0:5 соответствует наибольшей неуверенности классификатора в метке класса.

Мы использовали несколько более сложный метод оценки всех примеров, а затем выбирали b=2 примера, наиболее близких к 0:5 и выше, и b=2 примера, наиболее близких к 0:5 и ниже. Этот метод гарантирует, что не более половины примеров, выбранных на одной итерации, являются точными дубликатами (если только все примеры не оцениваются выше 0:5 или все ниже 0:5). Кроме того, есть некоторые свидетельства того, что обучение на парах примеров, находящихся по разные стороны границы принятия решения, является полезным [14].

4.4 Классификация с помощью вероятностного классификатора

Преимущество использования классификатора, который дает точные оценки P (Cjw), заключается в том, что при определенных предположениях теория принятия решений дает оптимальное правило для принятия решения о том, следует ли отнести пример к классу C ([27], p. 15). Пусть l_{ij} - это штраф или убыток, понесенный за принятие решения о классе i, когда истинным классом является j. (Пусть i; j = 1 для C, i; j = 2 для C) Тогда мы должны отнести пример к классу C именно тогда, когда:

$$l_{21} P(Cjw) + l_{22} (1 P(Cjw)) > l_{11} P(Cjw) + l_{12} (1 P(Cjw)) \qquad (10)$$

Например, если мы хотим получить минимальный процент ошибок (оба типа неправильных решений одинаково плохи), то соответствующие потери будут равны $l_{12} = l_{21} = 1$ и $l_{11} = l_{22} = 0$.

5 Эксперимент

Мы провели эксперимент, чтобы проверить, уменьшит ли выборка неопределенности количество помеченных данных, необходимых для обучения классификатора, по сравнению со случайной выборкой и выборкой релевантности. Использовались метод обучения и вероятностный классификатор из раздела 4. Классификаторы были обучены выполнять задачу категоризации текста на заголовках новостей.

	Обучение		Тест	
Категория	Номер	Частота.	Номер	Частота.
tickertalk	208	0.0007	40	0.0008
коробочный це облигации	314 470	0.0010 0.0015	42 60	0.0008 0.0012
nielsens	511	0.0016	87	0.0017
Бурма	510	0.0016	93	0.0018
дукакис	642	0.0020	107	0.0021
Ирландия	780	0.0024	117	0.0023
Квейл	786	0.0025	133	0.0026
бюджет	1176	0.0037	197	0.0038
заложники	1560	0.0049	228	0.0044

Таблица 1. 10 категорий, использованных в наших экспериментах, с указанием количества и частоты встречаемости на обучающем и тестовом наборах.

5.1 Набор данных

Заголовки 371 454 статей, появившихся в новостной ленте AP в период с 1988 по начало 1993 года, были случайным образом разделены на обучающий набор из 319 463 заголовков и тестовый набор из 51 991 заголовка. Заголовки были обработаны путем выделения текста нижним регистром и удаления знаков препинания. Границы слов определялись пробелами. Для минимизации вычислений использовались заголовки, а не полный текст статей.

Присваиваемые категории основывались на \keyword" из строки \keyword slug", присутствующей в каждом пункте AP ([28], p. 317). Ключевое слово - это строка длиной до 21 символа, указывающая на содержание элемента. Хотя ключевые слова должны быть идентичными только для обновляемых статей об одном и том же сюжете, на практике происходит значительное повторное использование ключевых слов и их частей из статьи в статью и из года в год, поэтому они имеют некоторые аспекты контролируемого словаря.

Мы определили категории названий AP в зависимости от того, встречались ли определенные подстроки в ключевом слове например. Например, следующие истории были отнесены к категории облигаций (ключевое слово выделено жирным шрифтом):

Сберегательные облигацииПродажи сберегательных облигаций упали после снижения ставки
СбереженияОблигации Казначейство объявило о снижении ставки по сбережениям на 2 процента СбереженияОблигацииЖертвам наводнения разрешено досрочно обналичить сберегательные облигации MesaBonds Mesa начнет обмен облигаций на 600 миллионов долларов в среду BondFirms Report: Облигационные фирмы Уолл-стрит запретят политические взносы Obit-Bond James Bond, Ornithologist, Gave Name To Fictional Agent 007
Люди-бонды Джулиан Бонд: Движению за права человека нужны личности, а не харизматичные лидеры

а эти - нет:

Избранный президент Клинтон играет в футбол
Bank-Failures Bank, S&L Failures at Seven-Year Low Taxes:SavingsBon
Taxes: Сберегательные облигации
КазначействоЗаимствования Казначейство смещает заимствования с долгосрочных облигаций MuniProbe Предложены более жесткие правила политических взносов для облигаций Muni

Категории, определенные таким образом, оказались несколько запутанными с семантической точки зрения. Джулиан Бонд и Джеймс Бонд не должны включаться в категорию сберегательных облигаций, а статьи о финансовых облигациях мы теряли, если ключевое слово было усечено, неправильно написано или подчеркивало какой-то другой аспект истории. Поэтому идеальная категоризация с помощью этих определений категорий невозможна. Мы не считаем это серьезной проблемой, поскольку нас интересует относительная, а не абсолютная эффективность методов категоризации. Более того, эти категории служат полезным тестом на устойчивость наших методов к ошибкам в обучающих данных.

10 категорий, которые мы определили, показаны в таблице 1 вместе с их частотой в обучающем и тестовом наборах. Категории были выбраны таким образом, чтобы иметь относительно низкую частоту, но при этом обеспечивать достаточное количество положительных примеров как в обучающем, так и в тестовом наборах.

5.2 Обучение

Начальный классификатор, необходимый для алгоритма выборки неопределенности (рис. 3), может быть получен из набора слов, предложенных учителем, подобно тому, как классификаторы создаются из текстов запросов пользователей в текстовых поисковых системах [29]. Чтобы избежать предвзятости экспериментаторов, мы использовали стартовую подвыборку из

положительных примеров категории, случайно выбранных из обучающего набора. При выборе признаков всегда использовались слова из этих трех примеров в дополнение к словам, выбранным, как описано в разделе 4.2.

В каждом запуске 3 начальных примера использовались для обучения начальной классификации, после чего выполнялось 249 итераций.

выборки неопределенности с размером подвыборки 4 были проведены, как описано в разделе 4.3. После того как каждая подвыборка была отобрана, метки ее категорий просматривались, и примеры добавлялись в набор помеченных примеров для использования в обучении следующего классификатора. Классификатор, полученный на каждой итерации, использовался для отбора примеров на следующей итерации, а также сохранялся для оценки. Чтобы изучить влияние начальной подвыборки на качество итоговой классификации, мы повторили этот процесс 10 раз для каждой категории, каждый раз с разной начальной подвыборкой из 3 положительных примеров.

Мы сравнили выборку неопределенности с выборкой релевантности и случайной выборкой. Выборка релевантности проводилась так же, как и выборка неопределенности, за исключением того, что были выбраны 4 примера с наибольшими значениями P (Cjw), а не 4 со значениями, близкими к 0,5.

\random" образцы фактически объединили начальные подвыборки из 3 положительных примеров с действительно случайными образцами различных размеров. Таким образом, были получены обучающие наборы следующих размеров:

3 6 10 20 40 80 160 320 640 1000 2500 4000 6000 8000 10000 15000 20000 30000 40000 50000 60000

70000 80000 ... (по 20000) ... 300000 319463

Большие наборы включали в себя меньшие. На основе каждого из этих наборов был сформирован классификатор с использованием тех же методов обучения, которые применялись для выборки неопределенности и релевантности, но классификатор использовался только для оценки, а не для управления выборкой. Из каждой из 10 начальных подвыборок для категории было сделано по два прогона, что дало в общей сложности 20 прогонов для каждой категории.

5.3 Оценка

Мы рассматривали каждую из 10 категорий как задачу бинарной классификации и оценивали классификаторы для каждой категории отдельно. Классификаторы оценивались путем применения их с минимальными параметрами потерь ($l_{12} = l_{21} = 1$ и $l_{11} = l_{22} = 0$) к 51 991 тестовому элементу и сравнения решений классификаторов с реальными метками категорий. Оценивались все классификаторы, обученные на случайных выборках. Оценивались классификаторы, сформированные в течение первых 10 итераций выборки неопределенности и релевантности, а также на каждой 5-й итерации после этого.

Для категоризации текста показатели эффективности recall и precision определяются следующим образом:

$$recall = \frac{\text{Количество членов категории в тестовом наборе, отнесенных к категории}}{\text{Количество членов категории в тестовом наборе}} \tag{11}$$

$$точность = \frac{\text{Количество членов тестового набора, отнесенных к категории}}{\text{Общее количество членов тестового набора, отнесенных к категории}} \tag{12}$$

При сравнении двух классификаторов желательно иметь единую меру эффективности. Ван Райсберген определил E-меру как комбинацию recall (R) и precision (P), удовлетворяющую определенным теоретическим свойствам ([30], pp. 168-176):

$$E = 1 - \frac{(^2 + 1)P\,R}{^2P + R} \tag{13}$$

Параметр находится в диапазоне от 0 до in nity и управляет относительным весом, придаваемым отзыву и точности. Значение 1 соответствует одинаковому весу отзыва и точности. Чтобы получить единую меру эффективности, где более высокие значения соответствуют лучшей эффективности, и где отзыв и точность имеют равное значение, мы определяем $F_{=1} = 1 - E_{=1}$.

6 Результаты

В таблице 2 для каждой категории приведены средние значения $F_{=1}$ для классификаторов, сформированных по выборке неопределенности, а также по выборке релевантности и на полном обучающем множестве. Мы также показываем эффективность, используя только 3 начальных примера и 7 случайно выбранных примеров. Это дает представление о качестве первоначального классификатора. Для всех категорий, кроме tickertalk, выборка неопределенности из 999 текстов привела к классификатору, значительно более эффективному, чем исходный классификатор или классификатор, сформированный на основе выборки релевантности из 999 текстов. Как правило, эффективность такой классификации была такой же или выше, чем у классификатора, обученного на всех 319 463 текстах. Классификаторы, обученные на случайной выборке из 1000 текстов, в большинстве случаев имели очень низкую эффективность.

На рисунке 2 показана зависимость эффективности от размера выборки для выборки неопределенности, выборки релевантности и случайной выборки. Представлены результаты по 9 категориям, без учета tickertalk, по которым ни одна стратегия не сработала.

Категория	3 + 996 не знаю.		3 + 7 рэнд.		3 + 996 rel.		3 + 319 460 полный	
	среднее	SD	средн ее	SD	среднее	SD	среднее	SD
tickertalk	.033	(.031)	.018	(.023)	.023	(.039)	.047	(.001)
коробочн ый це	.700	(.041)	.222	(.172)	.481	(.053)	.647	(.023)
облигации	.636	(.034)	.146	(.134)	.541	(.069)	.509	(.020)
nielsens	.801	(.016)	.291	(.218)	.567	(.132)	.741	(.022)
Бурма	.653	(.035)	.032	(.033)	.201	(.057)	.464	(.023)
дукакис	.136	(.046)	.101	(.075)	.035	(.021)	.163	(.015)
Ирландия	.416	(.041)	.050	(.033)	.170	(.038)	.288	(.030)
Квейл	.386	(.040)	.081	(.064)	.140	(.072)	.493	(.009)
бюджет	.290	(.039)	.058	(.046)	.141	(.029)	.235	(.005)
заложники	.477	(.021)	.068	(.042)	.177	(.039)	.498	(.003)

Таблица 2. Среднее и стандартное отклонение $F=1$ для обучения на начальных 3 примерах в сочетании с каждым из 996 примеров, отобранных без определенности, 7 случайных примеров, 996 примеров, отобранных по релевантности, или 319 460 оставшихся примеров. Средние значения приведены за 10 прогонов для выборки по неопределенности и релевантности и за 20 прогонов для случайной и полной выборки.

7 Обсуждение

Как показано на рисунке 2, эффективность классификации в целом увеличивается с ростом размера выборки при всех методах выборки, но быстрее при двух последовательных методах. Из последовательных методов выборка неопределенности значительно превосходит выборку релевантности. Эти результаты справедливы для категорий с сильно различающимися абсолютными уровнями эффективности.[4]

Превосходство выборки неопределенности над выборкой релевантности особенно заметно, поскольку низкая частота используемых категорий ограничивает опасность того, что выборка релевантности утонет в положительных примерах. Действительно, разница между выборкой неопределенности и выборкой релевантности ниже для менее частых категорий. Однако даже здесь выборка неопределенности лучше, как по более высокому среднему значению, так и по более низкому стандартному отклонению.

В большинстве случаев уровень эффективности, достигнутый при использовании случайной выборки только при 100 000 или более обучающих примеров, достигается при использовании выборки неопределенности на менее чем 1000 примеров, а обучение на 1000 случайно отобранных примерах дает значительно худшие результаты. В некоторых случаях для достижения заданного уровня эффективности требуется в 500 раз больше случайно отобранных примеров, чем примеров, отобранных с помощью выборки неопределенности.

Следует проявить некоторую осторожность при сравнении результатов для малых выборок неопределенности с результатами для больших случайных выборок. Для 6 из 10 категорий среднее значение $F=1$ для классификатора, обученного на выборке неопределенности из 999 примеров, фактически превышает значение, полученное при обучении на полном обучающем множестве из 319 463 примеров. Это означает, что какой-то аспект нашего обучения классификаторов неэффективно использует большие обучающие наборы. Наиболее вероятным злодеем является отбор признаков. Наш метод позволил получить несколько тысяч признаков при применении к

полный обучающий набор, и предыдущие работы показывают, что это слишком много [19].

Графики средней эффективности скрывают некоторую вариативность от прогона к прогону. Несколько стандартных отклонений, приведенных в таблице 2, составляют 10 % и более от средней эффективности, что означает некоторую непредсказуемость качества нальных классификаторов.

Одним из источников этой вариации были наши первоначальные подвыборки из 3 положительных примеров. Когда подвыборка была

плохо представляли категорию, начальный классификатор был неэффективен, и происходила значительная задержка, прежде чем начиналась выборка неопределенности для поиска дополнительных положительных примеров. Даже начальные классификаторы с одинаковой исходной эффективностью могли приводить к тому, что вначале искались разные части пространства примеров. Помимо влияния на раннее формирование классификатора, стартовые подвыборки оказывали влияние на эктивность за счет того, что мы делали их слова необходимыми характеристиками. Это видно по стандартным отклонениям для столбца Full таблицы 2, где, поскольку все классификаторы были обучены на одном и том же наборе примеров, единственным

Разница между прогонами для одной категории заключается в требуемых характеристиках, предоставляемых начальными примерами.

Вторым источником вариаций были колебания в качестве последовательных классификаторов. Процесс выборки неопределенности по своей сути является исследовательским, и недостатки в классификаторе, полученном на одной итерации, приводят к выбору компенсирующих примеров на последующих итерациях.

8 Будущая работа

Наши результаты показывают, что эффективные текстовые классификаторы могут быть созданы путем получения большого количества немеченых данных и маркировки лишь небольшой их части. Хотя мы тестировали выборку неопределенности на задаче категоризации текста, ее можно с равным успехом применить к любой задаче классификации.

[4] *Различия в абсолютных уровнях эффективности вполне ожидаемы: некоторые категории просто сложнее других, а некоторые из наших определений категорий были особенно шумными.*

Поиск текстов - очевидное приложение, но при этом необходимо учитывать компромисс между поиском текстов, наиболее полезных для пользователя, и текстов, на основе которых система сможет извлечь наибольшую пользу [31]. В приложениях для сортировки, маршрутизации и распространения информации этот компромисс менее актуален, поскольку затраты на оценку нерелевантных примеров могут быть амортизированы за более длительный период работы.

Выборка неопределенности также должна принести пользу подходам к задачам обработки естественного языка, основанным на классификации. В ряде проектов были аннотированы или аннотируются большие корпорации для поддержки обучения статистических методов для этих задач. Наши результаты показывают, что сбор огромных немаркированных корпораций и использование выборки неопределенности для аннотирования небольшого подмножества для каждой задачи может быть дешевле и не менее эффективным. Кроме обработки текстов, могут быть полезны и другие области, где доступны большие наборы данных.

Еще предстоит ответить на многие вопросы, связанные с выборкой неопределенности. Наиболее важные практические вопросы связаны с тем, как учитель узнает, когда нужно остановиться, и как сформировать окончательный классификатор для использования после этого. Оценки эффективности классификатора позволили бы учителю отслеживать прогресс, но для получения таких оценок на основе неслучайной выборки потребуются новые методы. Оценки эффективности также помогут выбрать классификатор для использования из классификаторов, сформированных на последних нескольких итерациях. В качестве альтернативы можно использовать методы стабилизации к о л е б а н и й от итерации к итерации.

Можно изучить множество расширений и улучшений выборки неопределенности. Нам необходимо определить связь между размером подвыборки и ее эффективностью, поскольку большие подвыборки требуют меньше вычислений. Также кажется вероятным, что размер подвыборки можно увеличить, если уменьшить избыточность внутри подвыборок. Другие способы повышения эффективности включают использование менее точного, но более эффективно обученного классификатора во время выборки [32], а также выбор первых примеров, удовлетворяющих порогу неопределенности, а не самых неопределенных примеров. Также представляет интерес одновременное обучение классификаторов для нескольких классов.

В текущей формулировке выборка неопределенности требует, чтобы лежащий в основе алгоритм обучения создавал разумные классификаторы даже на очень маленьких обучающих наборах. Это означает, что нам пришлось повозиться с выбором признаков и оценкой параметров, чтобы избежать патологического поведения на малых выборках. В настоящее время мы изучаем варианты выборки неопределенности, которые будут более устойчивы к проблемам обучения классификаторов.

9 Резюме

Текст стоит дешево, но информация, в виде знания о том, к каким классам относится текст, стоит дорого. Автоматическая классификация текста может предоставить эту информацию по низкой цене, но сами классификаторы должны быть созданы с помощью дорогостоящего человеческого труда или обучены на основе текстов, которые были классифицированы вручную. Мы продемонстрировали, что выборка неопределенности может резко сократить объем текста, который необходимо вручную маркировать для создания эффективного классификатора. Выборка неопределенности имеет потенциальное применение в различных задачах обработки текстов, а также в других областях, где доступны большие объемы неклассифицированных данных.

10 Благодарности

Мы благодарим Джейсона Кэтлетта, Уильяма Коэна, Эйлин Фитцпатрик, Йоава Фройнда, Тревора Хасти, Роберта Шапира и Себастьяна Сеунга за советы и полезные комментарии к этой работе, а также Кена Черча за предоставленные инструменты обработки текста и помощь в работе с ними.

Ссылки

1. *P. J. Hayes. Интеллектуальная обработка больших объемов текста с использованием неглубоких, специфичных для конкретной области методов. In Paul. S. Jacobs, editor, Text-Based Intelligent Systems: Current Research in Text Analysis, Infor- mation Extraction, and Retrieval, pages 227{241. Lawrence Erlbaum, Hillsdale, NJ, 1992.*
2. *P. Biebricher, N. Fuhr, G. Lustig, M. Schwantner, and G. Knorz. Система автоматического индексирования AIR/PHYS|от исследования к применению. В Трудах SIGIR-88, страницы 333|342, 1988.*
3. *У. Г. Кокран. Sampling Techniques. John Wiley & Sons, New York, 3-е издание, 1977.*
4. *Г. Солтон и К. Бакли. Улучшение эффективности поиска с помощью обратной связи по релевантности. Журнал Американского общества информационной науки, 41(4):288{297, 1990.*
5. *В. А. Гейл, К. В. Черч и Д. Яровски. Метод деамбигуации смыслов слов в большом корпусе. Computers and the Humanities, 26:415{439, 1993.*
6. *Б. К. Гхош. Краткая история последовательного анализа. В B. K. Ghosh и P. K. Sen, редакторы, Handbook of Sequential Analysis, глава 1, страницы 1{19. Marcel Dekker, New York, 1991.*
7. *Д. Англуин. Запросы и обучение концепциям. Машинное обучение, 2:319{342, 1988.*

8. М. Плутовски и Х. Уайт. Выбор кратких обучающих наборов из чистых данных. IEEE Transactions on Neural Networks, 4(2):305{318, март 1993.

9. Д. Кон, Л. Атлас и Р. Ладнер. Улучшение обобщения с помощью самонаправленного обучения, 1992. Появится в журнале Machine Learning.

10. D. J. C. MacKay. Система доказательств в применении к классификационным сетям. Нейронные вычисления, 4:720{736, 1992.

11. Х. С. Сон, М. Оппер и Х. Сомполински. Запрос по комитету. В материалах Пятого ежегодного семинара ACM по теории вычислительного обучения, страницы 287{294, 1992.

12. Т. М. Митчелл. Обобщение как поиск. Arti cial Intelligence, 18:203{226, 1982.

13. Й. Фройнд, Х. С. Сюн, Э. Шамир и Н. Тишби. Информация, предсказание и запрос по комитету. In Advances in Neural Informations Processing Systems 5, San Mateo, CA, 1992. Morgan Kaufmann.

14. J. Hwang, J. J. Choi, S. Oh, and R. J. Marks II. Обучение на основе запросов, применяемое к частично обученным многослойным перцептронам. IEEE Transactions on Neural Networks, 2(1):131{136, January 1991.

15. Д. Т. Дэвис и Дж. Хванг. Обучение фокусу внимания с помощью выбора данных о граничных областях. In In- ternational Joint Conference on Neural Networks, pages I{676 to I{681, Baltimore, MD, June 7{11 1992.

16. П. Э. Харт. Сокращенное правило ближайшего соседа. IEEE Transactions on Information Theory, IT-14:515{516, May 1968.

17. R. E. Utgo . Улучшенное обучение с помощью инкрементального обучения. Шестой международный семинар по машинному обучению, страницы 362{365, 1989.

18. Н. Фур. Модели для поиска с вероятностным индексированием. Обработка информации и управление, 25(1):55{72, 1989.

19. Д. Д. Льюис. Оценка фразовых и кластерных представлений в задаче категоризации текста. В Трудах SIGIR-92, страницы 37{50, 1992.

20. М. Э. Марон. Автоматическое индексирование: Экспериментальное исследование. Journal of the Association for Com- puting Machinery, 8:404{417, 1961.

21. У. С. Купер. Некоторые несоответствия и неправильные термины в вероятностном информационном поиске. В Трудах SIGIR-91, страницы 57{61, 1991.

22. П. Маккаллах и Дж. А. Нелдер. Обобщенные линейные модели. Chapman & Hall, London, 2nd edition, 1989.

23. У. С. Купер, Ф. К. Гей и Д. П. Дабни. Вероятностный поиск на основе поэтапной логистической регрессии. В Трудах SIGIR-92, страницы 198{210, 1992.

24. Н. Фур и У. Пфайфер. Объединение подходов, ориентированных на модели и описания, для вероятностного индексирования. В Трудах SIGIR-91, страницы 46{56, 1991.

25. С. Робертсон и Дж. Бови. Статистические проблемы применения вероятностных моделей для поиска информации. Отчет 5739, Британская библиотека, Лондон, 1982.

26. У. А. Гейл и К. У. Черч. Плохие оценки контекста хуже, чем никакие. На семинаре "Речь и естественный язык", стр. 283{287, Сан-Матео, Калифорния, июнь 1990 г. DARPA, Morgan Kaufmann.

27. R. O. Duda and P. E. Hart. Pattern Classi cation and Scene Analysis. Wiley-Interscience, New York, 1973.

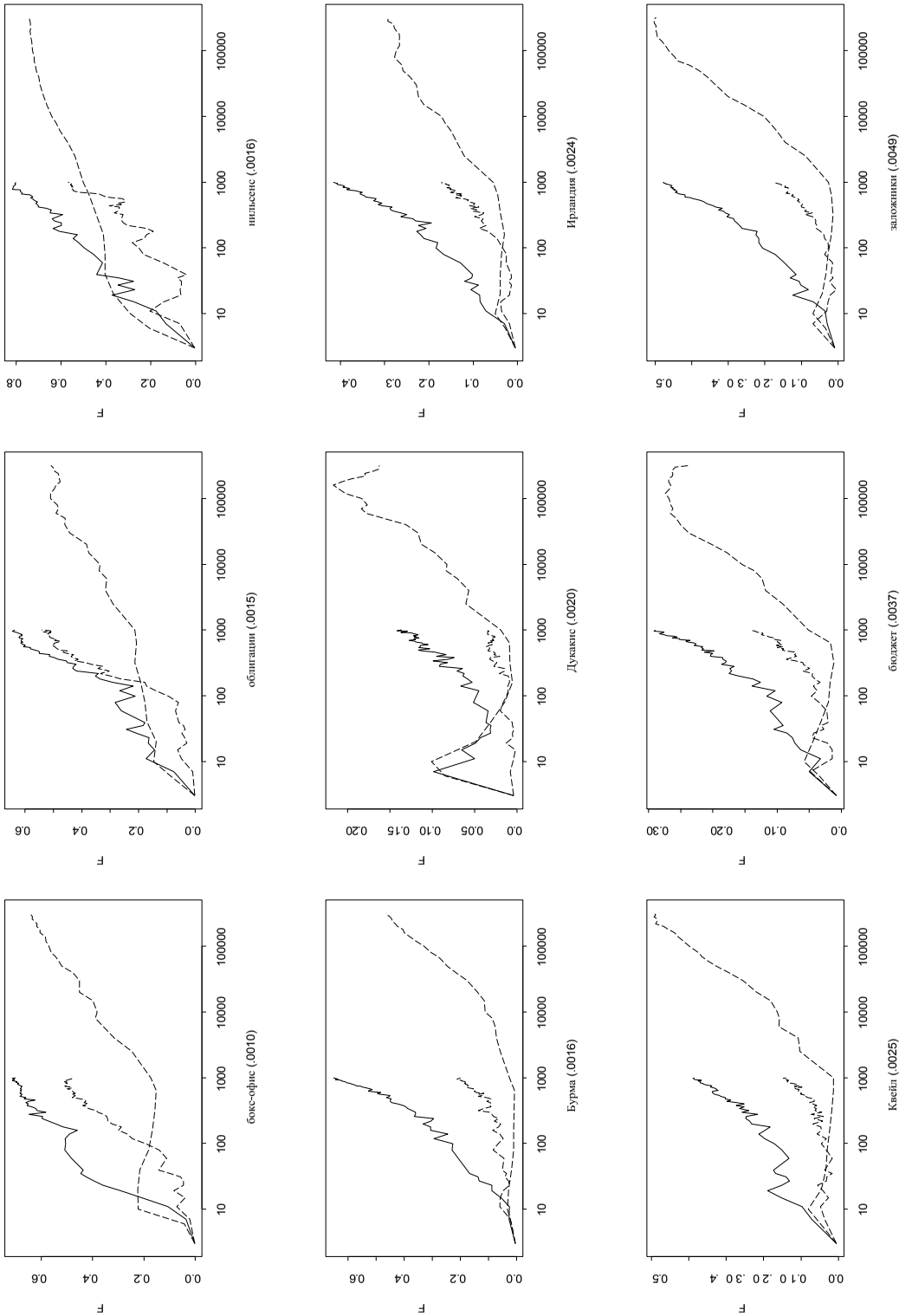
28. Н. Голдштейн, редактор. The Associated Press Stylebook and Libel Manual. Addison-Wesley, Reading, MA, 1992.

29. В. Б. Крофт и Д. Дж. Харпер. Использование вероятностных моделей поиска документов без обратной связи по релевантности. Journal of Documentation, 35(4):285{295, 1979.

30. C. J. van Rijsbergen. Information Retrieval. Butterworths, London, второе издание, 1979.

31. А. Букштейн. Информационный поиск: Последовательный процесс обучения. Журнал Американского общества информационной науки, 34:331{342, сентябрь 1983.

32. Дэвид Д. Льюис и Джейсон Кэтлетт. Гетерогенная выборка неопределенности для контролируемого обучения. В материалах Одиннадцатой международной конференции по машинному обучению, 1994. To appear.



Рисуно 2. Среднее значения для текстовых классификаторов, обученных на неопределенности (сплошная линия), к значение $F=1$ релевантности (пунктирная линия) и случайные (пунктирная линия) выборки названий из корпуса AP. Средние значения за 10 прогонов для неопределенности и релевантности

выборки, и более 20 прогонов для случайной выборки. Результаты показаны для 9 категорий. В скобках указана частота встречаемости категорий в обучающем наборе.