Кемеровский государственный университет

Томский государственный университет

Кемеровский научный центр Сибирского отделения РАН

Филиал Кемеровского государственного университета

в г. Анжеро-Судженске

НАУЧНОЕ ТВОРЧЕСТВО МОЛОДЕЖИ

Материалы X Всероссийской научно-практической конференции

21 – 22 апреля 2006 г.

Часть 1

Издательство Томского университета

2006

ББК 74+72
Н76

Научное творчество молодежи: Материалы X Всероссийской Н76 научно-практической конференции (21-22 апреля 2006 г.) Ч. 1. – Томск: Изд-во Том. ун-та, 2006. – 192 с.

ISBN 5-7511-2015-3

В ч. 1 материалов конференции вошли тезисы докладов по секциям «Естествознание, биология, медицина», «Информатика», «Математические методы в технических приложениях» и «Прикладная математика и математическое моделирование».

ББК 74+72

Конференция организована при поддержке Российского фонда фундаментальных исследований (проект № 06-06-85031).

Руководитель проекта – д-р техн. наук, проф. Е.В. Глухова.

Конференция проводится в рамках мероприятий, посвященных 15-летию филиала КемГУ в г. Анжеро-Судженске.

ISBN 5-7511-2015-3

КЛАССИФИКАЦИЯ ТЕКСТОВ ПО АВТОРСТВУ С ПОМОЩЬЮ МЕТОДА ХМЕЛЕВА И ЕГО МОДИФИКАЦИЙ

В.В. Поддубный, О.Г. Шевелев
Томский государственный университет

Классификация текстов по количественным признакам производится в настоящее время с помощью самых различных методов: нейронных сетей, метода опорных векторов (SVM), метода сжатия данных, дискриминантного анализа и других. Одним из наиболее эффективных является метод Хмелева [1, 2], который позволяет с высокой точностью классифицировать тексты по авторству. В данной работе предлагаются модификации метода и результаты исследований по определению качества классификации с их помощью.

Сутью метода является подсчет и обработка парных сочетаний элементов текста (сочетаний букв [1] или сочетаний грамматических классов [2]). Классификации предшествует обучение системы. Обучение производится на текстах заданного множества авторов. Для каждого автора подсчитывается матрица-эталон употреблений всех пар рассматриваемых элементов в его текстах. При распознавании авторства произвольного текста подсчитывается аналогичная матрица и сравнивается со всеми имеющимися матрицами-эталонами. Автор, обладающий наиболее похожей матрицей-эталоном, выбирается в качестве автора рассматриваемого текста.

Несмотря на то, что в работе [1] о матрицах переходов говорится как о реализациях цепей Маркова, свойство марковости, определяемое уравнением Чепмена-Колмогорова [3] для вероятностей переходов, нигде не проверяется и не используется. Поэтому распознавание авторства с помощью таких матриц, на наш взгляд, корректнее называть распознаванием по частотам переходов.

Методы, работающие на основе матриц частот переходов, могут варьироваться в зависимости о того, какие именно переходы подсчитываются (букв, слов, предложений, любых или с определенными свойствами) и какая мера используется для сравнения матриц. В работах [1, 2] в качестве меры сравнения использовалась величина

$$L = \sum_{i=1}^{k}\sum_{j=1}^{k} m_{1ij} \cdot \ln\left(\frac{m_{1ij}}{n_{1i}} \Big/ \frac{m_{2ij}}{n_{2i}} \right), \qquad (1)$$

где m_{1ij} – число переходов из i элемента в j в анализируемом тексте, n_{1i} – общее число переходов из i-го элемента, m_{2ij}, n_{2i} – аналогичные числа для матрицы того автора, с которым производится сравнение, k – число элементов (если подсчитываются буквосочетания, то k = 32). Значение L тем меньше по модулю, чем меньше различие между матрицами. Меру L будем называть «мерой Хмелева». Легко видеть, что эта мера с точностью до постоянного множителя совпадает с информационной мерой расхождения условных распределений (условных частот), известной в статистике как направленная дивергенция Кульбака [4, 5]:

$$I = 2 \cdot \sum_{i=1}^{k}\sum_{j=1}^{k} \left(\frac{m_{1ij}}{n_1} \right) \cdot \ln\left(\frac{m_{1ij}}{n_{1i}} \Big/ \frac{m_{2ij}}{n_{2i}} \right). \qquad (2)$$

Будем называть эту величину «условной» мерой Кульбака. Аналогичным образом может быть построена «безусловная» мера Кульбака, являющаяся информационной мерой расхождения безусловных распределений (частот пар):

$$I = 2 \cdot \sum_{i=1}^{k} \sum_{j=1}^{k} \left(\frac{m_{1ij}}{n_1} \right) \cdot \ln \left(\frac{m_{1ij}}{n_1} \Big/ \frac{m_{2ij}}{n_2} \right), \qquad (3)$$

где $n_1 = \sum_{i=1}^{k} n_{1i}$, $n_2 = \sum_{i=1}^{k} n_{2i}$. Очевидно, «безусловная» мера Хмелева, построенная по формуле (1) с использованием частот пар m_{1ij}/n_1 и m_{2ij}/n_2 вместо условных частот m_{1ij}/n_{1i} и m_{2ij}/n_{2i}, также с точностью до постоянного множителя совпадает с «безусловной» мерой Кульбака (3).

Меры, представленные формулами (1), (2) и (3), «направлены» от матрицы анализируемого текста к матрице-эталону (усреднение по анализируемому тексту). Возможны и другие варианты – направленность на анализируемый текст (усреднение по матрице-эталону), симметричная мера (сумма мер в одну и в другую сторону пополам).

Значения логарифмов отношений частот под суммой в формулах (1) и (2) могут быть как положительными, так и отрицательными. Общая сумма может то уменьшаться, то увеличиваться, в зависимости от значений строк матрицы. Поэтому различие между матрицами не будет расти постоянно. Для обеспечения более стабильных мер возьмем каждое слагаемое по модулю и получим еще две модификации меры. Назовем их модульными мерами Хмелева и Кульбака соответственно.

Другим вариантом сравнения матриц является подсчет статистики хи-квадрат, значение которой также может выступать в качестве меры различия распределений [6]:

$$\chi^2 = n_1 n_2 \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{1}{m_{1ij} + m_{2ij}} \cdot \left(\frac{m_{1ij}}{n_1} - \frac{m_{2ij}}{n_2} \right)^2 . \qquad (4)$$

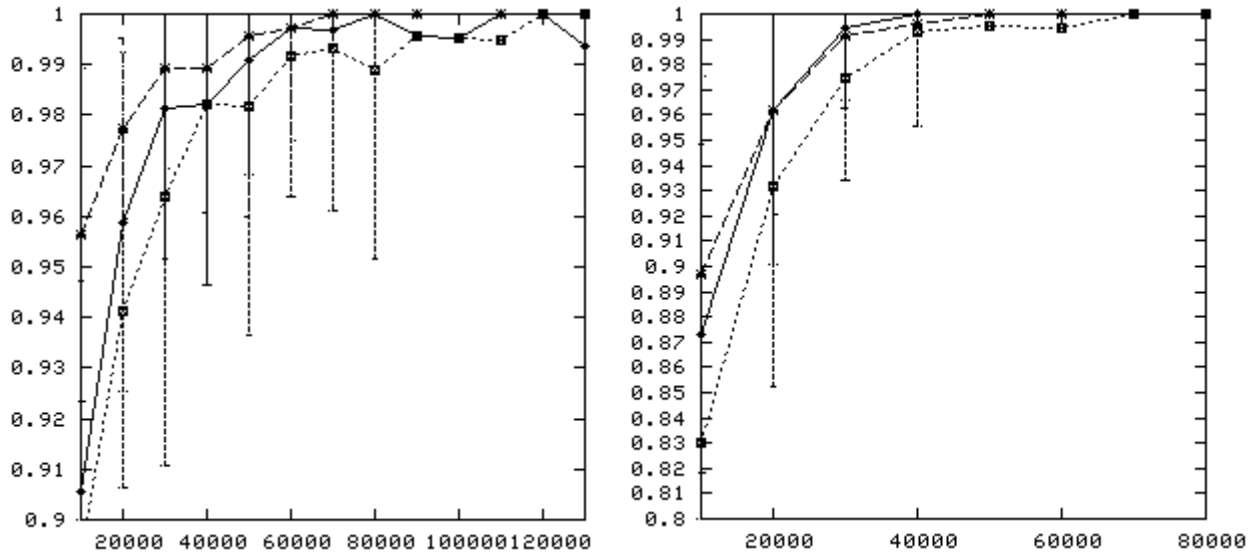
Эта мера построена для двумерных распределений частот переходов исследуемых выборок (частот пар). Мера хи-квадрат в отличие от вышерассмотренных мер является ненаправленной. Более того, как видно из формулы, данная мера, как и «безусловная» направленная дивергенция Кульбака (3), не использует специфической информации о матрице частот переходов, в отличие от меры Хмелева (1) и «условной» направленной дивергенции Кульбака (2), в которых подсчитывается сумма частот по строкам матрицы (n_{1i}, n_{2i} – общее число переходов из *i*-го элемента). Данное обстоятельство, с одной стороны, говорит о недостатке «безусловных» мер, так как, в отличие от меры Хмелева (вообще «условных» мер), в ходе классификации используется меньше информации о данных, с другой – о достоинстве. «Безусловная» модификация (3), (4) «условных» мер для сравнения с их помощью одномерных распределений позволяет использовать для классификации текстов не только частоты переходов, но и другие возможные характеристики.

Заметим, что «безусловная» направленная дивергенция Кульбака (3) (а, следовательно, и «безусловная» мера Хмелева) и статистика хи-квадрат (4) при верной нулевой гипотезе (равенстве генеральных частот переходов для обеих выборок) асимптотически распределены по одному и тому же закону χ^2 с $(k-1)^2$ степенями свободы [4].

Для оценки качества работы метода Хмелева при использовании вышеуказанных мер проведем ряд экспериментов по классификации различных текстов по авторству. Рассмотрим два набора текстов: художественные тексты 10 авторов (русские классики и современники) и газетные статьи 10 журналистов (из 4 томских газет). Классификацию проведем на основе частот появления пар букв (всего 1024 признака). Рассмотрим различные наборы данных, в каждом из которых тексты разбиты на фрагменты определенной длины. Качество классификации p_{ra} (right answers) в ходе каждого эксперимента будем оценивать по частоте правильно классифицированных фрагментов на тестовой выборке по методу *k*-подмножеств. Суть метода состоит в разделении исходных данных на *k* равных частей и запуске алгоритма (обучении и тестировании) *k* раз, причем в ходе каждого запуска (*k*-1) частей участвует в обучении, одна – в тестировании, а тестовая часть постоянно меняется. Возьмем *k* равным

10. Такой величины вполне хватает для оценки качества, и при этом подмножества не будут слишком малы.

В качестве результатов тестирования возьмем среднюю частоту $\overline{p_{ra}}$ правильных классификаций (среднее качество), полученную как среднеарифметическое каждого из k запусков алгоритма, и границы 95% интерквантильного интервала, задающие разброс частот. Для нормального распределения (а гипотеза о нормальности по критериям Колмогорова и Шапиро-Уилка для большей части исследуемых данных не отвергается) такие границы определяются интервалом $\overline{p_{ra}} \pm 2\sigma$. Для тех данных, для которых гипотеза о нормальности отвергается (это происходит, когда частоты слишком близко подходят к 1), в качестве границ частот берется 90% эмпирический интерквантильный интервал (в этот интервал попадает 9/10 всех частот p_{ra}, полученных в ходе тестирования по методу k-подмножеств, k = 10).



а) *б)*

Рис. 1. Классификация при использовании мер Хмелева (*), хи-квадрат (•), модульной меры Кульбака (□): а) художественные тексты, б) газетные статьи

Как видно из графиков на рис. 1, меры Хмелева, хи-квадрат, модульная мера Кульбака работают хорошо и примерно одинаково (нет значимого отличия частот). Качество классификации растет в среднем с увеличением объемов фрагментов, но начиная с критического значения (40000-60000 символов) стабилизируется и колеблется около 100%. Мера Кульбака и модульная мера Хмелева, не представленные на графике, работают плохо. Качество классификации с использованием меры Кульбака колеблется около 40-50%, с использованием модульной меры Хмелева – в некоторых случаях достигает значения немодульной меры, в некоторых – опускается заметно ниже.

ЛИТЕРАТУРА

1. Хмелев Д.В. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник МГУ. – Сер. 9: Филология. – 2000. – № 2. – С. 115-126.
2. Кукушкина О.В., Поликарпов А.А., Хмелёв Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. – 2001. – Т. 37. – Вып. 2. – С. 96-109.
3. Кемени Дж., Снелл Дж. Конечные цепи Маркова. – М.: Наука, 1982.
4. Закс Л. Статистическое оценивание. - М.: Статистика, 1976. – 600 с.
5. Кульбак С. Теория информации и статистика. – М.: Наука, 1967.
6. Крамер Г. Математические методы статистики. – М.: Мир, 1976. – 648 с.