УДК 519.765

Т. Г. Суровцова, С. П. Чистяков

О ПОСТРОЕНИИ СТАТИСТИЧЕСКИХ КРИТЕРИЕВ ДЛЯ АТРИБУЦИИ АВТОРСТВА ЛИТЕРАТУРНЫХ ТЕКСТОВ

Введение. Задача автоматической классификации текстов имеет большое практическое значение. Процедуры данной классификации применяются при обработке информационных потоков, таких как электронная почта и новости, рекламные объявления, создания каталогов в Интернете, при автоматическом реферировании и аннотировании. Близкой к ней (но имеющей и принципиальные отличия) является задача атрибуции литературных текстов, а именно, отнесение произведения к конкретным жанру, стилю, времени написания и, вероятно наиболее значимой среди них, определению авторства произведения. Отметим, что автоматическое установление авторства письменных текстов, помимо литературоведения, имеет важное значение в сфере безопасности и при защите авторских прав, уголовном и гражданском делопроизводстве, криминалистике.

Для решения задачи атрибуции авторства анонимных или псевдонимных литературных текстов широко применяются методы, основанные на статистическом анализе количественных характеристик текстов – лингвостатистических параметров (см., например, [1–6]). Многие методики атрибуции авторства применяют аппарат статистической проверки гипотез. Наиболее часто для этой цели использовались критерии Стьюдента и Колмогорова [2, 4, 6]. Однако они дают возможность проверить гипотезу об авторстве только по одному параметру. Для того чтобы сравнить анализируемые тексты по нескольким параметрам, необходимо применить эти критерии к каждому лингвостатистическому параметру и затем синтезировать результаты этих проверок. На последнем этапе информация о статистической значимости результатов отдельных проверок, как правило, теряется. Поэтому желательно иметь статистический критерий, статистика которого зависела бы от всех имеющихся лингвостатистических параметров.

В статье предложен подход, применимый как к классификации, так и к атрибуции текстов, основанный на индуктивном построении классификаторов предложений. В задачах классификации текст относится к тому классу, к которому отнесено большинство из составляющих его предложений. На основе полученного классификатора разработана (и апробирована на ряде литературных текстов) процедура построения статистического критерия проверки нулевой гипотезы о том, что данный текст принадлежит

Суровцова Татьяна Геннадьевна – кандидат технических наук, преподаватель кафедры теории вероятностей и анализа данных математического факультета Петрозаводского государственного университета. Количество опубликованных работ: 10. Научное направление: автоматическая обработка текстов. E-mail: tsurovceva@psu.karelia.ru.

Чистяков Сергей Павлович – кандидат технических наук, младший научный сотрудник Института прикладных математических исследований Карельского научного центра РАН. Количество опубликованных работ: 43. Научные направления: прикладная статистика, распознавание образов. E-mail: chistiakov@krc.karelia.ru.

© Т. Г. Суровцова, С. П. Чистяков, 2009

некоторому конкретному автору. Статистикой критерия является количество предложений анализируемого текста, отнесенных классификатором не к этому автору.

Построение статистического критерия. Представим способ построения статистического критерия на основе некоторого классификатора.

Пусть $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ – некоторый вектор признаков и $\mathrm{X} = \mathbf{Dom}(\mathbf{X})$ – множество его возможных значений. Обозначим через Y классовый признак с множеством возможных значений $D = \{0, 1, \ldots, k-1\},\ k \geqslant 2$. Предположим, что определен некоторый классификатор $f : \mathrm{X} \to D$. Будем считать также, что существует неизвестное совместное распределение $P(\mathbf{x}, y)$ признаков $X_1, X_2, \ldots, X_n, Y$. Обозначим $P_i(\mathbf{x}) = P(\mathbf{x}|Y = i)$ условное распределение вектора признаков $\mathbf{X}$ при данном $Y = i$. Пусть $\mathcal{D}_0 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ – случайная выборка из одного из распределений $P_i(\mathbf{x}),\ i = 0, 1, \ldots, k-1$. Через H_0 обозначим нулевую гипотезу: выборка $\mathcal{D}_0$ является выборкой из распределения $P_0(\mathbf{x})$. Альтернативная гипотеза H_1 : $\mathcal{D}_0$ есть выборка из одного из распределений $P_i(\mathbf{x}),\ i \neq 0$. Предположим, что $\mathbf{X}$ – случайный вектор признаков из распределения $P_0(\mathbf{x})$. Через p_0 обозначим вероятность ошибочной классификации $\mathbf{X}$, т. е. $p_0 = \mathbf{P}\{f(\mathbf{X}) \neq 0|H_0\}$, и пусть N_f – количество элементов $\mathbf{x}$ выборки $\mathcal{D}_0$ таких, что $f(\mathbf{x}) \neq 0$. Предположим, нулевая гипотеза H_0 верна; тогда случайная величина N_f распределена по биномиальному закону с параметрами N и p_0, т. е.

$$\mathbf{P}(N_f = k) = C_N^k {p_0}^k (1 - p_0)^{N-k},\ k = 0, 1, \ldots, N\ .$$

Естественно отвергнуть H_0, если N_f «намного» больше, чем можно было бы ожидать в случае истинности нулевой гипотезы H_0.

Определим теперь статистический критерий проверки нулевой гипотезы H_0 против альтернативы H_1. Пусть α – некоторое число такое, что $0 < \alpha < 1$. Через N_0 обозначим минимальное число такое, что

$$\sum_{k=N_0}^{N} C_N^k p_0^k (1 - p_0)^{N-k} \leqslant \alpha. \tag{1}$$

Тогда критерий для проверки нулевой гипотезы H_0 определяется критической областью $\{\mathbf{x} : N_f \geqslant N_0\}$, т. е. если $N_f \geqslant N_0$, гипотеза H_0 отвергается. Из (1) следует, что уровень значимости данного критерия не больше α. Так как статистика критерия N_f дискретна, не всегда можно построить критерий с уровнем значимости, в точности равным α. Тем не менее можно (как обычно) использовать p-значения статистики критерия N_f.

Рассмотрим реализацию описанного выше подхода в задаче атрибуции авторства. Пусть $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ – обучающая выборка, где $\mathbf{x}_i,\ i = 1, 2, \ldots, l$, является вектором лингвостатистических параметров (признаков), описывающих некоторое предложение, и $y_i \in D$ представляет собой метку, указывающую на автора предложения. Построим классификатор f посредством индукции по обучающей выборке $\mathcal{D}$ и оценим значение p_0 по тестовой выборке (или методом кросс-проверки). Пусть $\mathcal{D}_0 = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ – множество векторов признаков, описывающее анализируемый литературный текст (состоящий из N предложений), имеющий спорное авторство. По выборке $\mathcal{D}_0$ вычисляются статистика критерия N_f и соответствующее p-значение (в качестве p_0 используется его оценка). Проверка нулевой гипотезы H_0: «текст, описываемый $\mathcal{D}_0$, был написан автором с меткой 0» осуществляется сравнением полученного p-значения с выбранным уровнем значимости. Для реализации на практике описанного выше подхода необходимо определить некоторое множество лингвостатистических

параметров $X_1, X_2, \ldots, X_n$ и построить классификатор f по обучающей выборке $\mathcal{D}$. Эти моменты рассматриваются ниже. Необходимо также заметить, что построенный классификатор может быть использован для автоматической классификации текстов (например, с помощью метода голосования), но рассмотрение данного вопроса не входит в задачу настоящей статьи.

Лингвостатистические признаки. Реализация описанного выше подхода требует выбора некоторого набора лингвостатистических признаков, каждый из которых представляет собой некоторую характеристику предложения. Набор квантитативных признаков, которые могут быть установлены для любого литературного произведения, достаточно разнообразен. Нам необходимо было получить достаточно полное и в то же время не содержащее излишней информации описание каждого предложения текста. На основе предыдущих исследований и собственного опыта было выбрано 20 признаков, составляющих четыре группы: описывающие расположение частей речи на различных позициях предложения [2]; легко рассчитываемые квантитативные признаки [1, 2]; описывающие синтаксическую структуру предложения [4]; описывающие части речи, которые входят в предложение [1, 2].

Полный список использованных нами признаков следующий: 1) часть речи в первой позиции предложения; 2) часть речи во второй позиции предложения; 3) часть речи в третьей позиции предложения; 4) часть речи в третьей с конца позиции предложения; 5) часть речи в предпоследней позиции предложения; 6) часть речи в последней позиции предложения; 7) средняя длина слова в буквах в предложении; 8) количество слов в предложении; 9) тип предложения; 10) цель высказывания для простых и сложных предложений; 11) модальность предложения для простых и сложных предложений; 12) наличие главных членов для простого предложения; 13) способ соединения частей сложного предложения; 14) относительное количество глаголов в предложении; 15) относительное количество прилагательных в предложении; 16) относительное количество существительных в предложении; 17) относительное количество предлогов в предложении; 18) относительное количество союзов в предложении; 19) наличие причастий в предложении; 20) относительное количество частиц в предложении.

Описания морфологических и синтаксических признаков текстов в соответствии с грамматикой русского языка были получены с помощью экспертной системы, входящей в информационно-поисковый программный комплекс «СМАЛТ» [7]. Так как выбранный нами тип классификатора (описанный ниже) предполагает, что все признаки обучающей выборки измерены в номинальной шкале, а среди выбранных лингвостатистических признаков присутствуют признаки, измеренные в интервальной шкале (например, средняя длина слова в буквах), эти признаки предварительно были подвергнуты дискретизации. В результате область возможных значений непрерывного признака разбивалась на совокупность дизъюнктных интервалов таким образом, чтобы различие распределений классового признака для любой пары смежных интервалов было статистически значимо (использовался критерий однородности χ^2 при уровне значимости 0.01). Затем каждый интервал интерпретировался как одно значение нового номинального признака.

Системы правил и классификаторы. Кратко опишем системы правил, использованные нами для построения классификаторов. Отметим, что хотя эти системы в распознавании образов занимают скромное место, в ряде областей, где важно понимание причин, на основе которых принимается решение (и к которым, безусловно, относится проблема атрибуции литературных работ), их применение вполне оправдано.

Введем такие обозначения. Пусть $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ – некоторый вектор

номинальных признаков и $\mathrm{X} = \mathrm{X}_1 \times \mathrm{X}_2 \times \ldots \times \mathrm{X}_n$, где $\mathrm{X}_i = \{x_{i1}, x_{i2}, \ldots, x_{ir_i}\}$, $i = 1, 2, \ldots, n$, – множество возможных значений признака X_i, а также Y – классовый признак с множеством возможных значений $D = \{0, 1, \ldots, k-1\}$, $k \geqslant 2$; $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ – обучающая выборка. Рассмотрим множества правил вида

$$\text{“ЕСЛИ } \langle \textit{предпосылка} \rangle \text{ ТО } \langle \textit{следствие} \rangle \langle \textit{с весом } w \rangle\text{”} . \tag{2}$$

В (2) предпосылка C имеет вид

$$C = \{X_{\alpha_1} = x_{\alpha_1\beta_1}\} \cap \{X_{\alpha_2} = x_{\alpha_2\beta_2}\} \cap \ldots \cap \{X_{\alpha_r} = x_{\alpha_r\beta_r}\}$$

и следствие

$$C_i^* = \{Y = i\},\ i \in D.$$

Вес $w \in (0, 1)$ является мерой влияния предпосылки правила на следствие. Правила такого вида обозначим $C \Rightarrow C_i^*\ \langle w \rangle$. Пусть $C_1 \Rightarrow C_i^*\ \langle w_1 \rangle$ и $C_2 \Rightarrow C_i^*\ \langle w_2 \rangle$ – некоторые правила с одним и тем же следствием C_i^*. Синтез весов этих правил основан на такой функции комбинации весов [8]:

$$w_1 \oplus w_2 = \frac{w_1 w_2}{w_1 w_2 + (1 - w_1)(1 - w_2)}.$$

Пусть $\mathcal{R}$ – некоторое множество правил. Тогда для любой предпосылки C и заключения $C_i^* = \{Y = i\},\ i \in D$ можно вычислить следующий *композиционный вес*:

$$W(C_i^* | \, C, \mathcal{R}) = \bigoplus_{\alpha} w_\alpha,$$

здесь функция комбинации весов $\bigoplus$ применяется к весам w_α всех правил $C' \Rightarrow C_i^*\ \langle w_\alpha \rangle$, содержащихся в $\mathcal{R}$ таких, что предпосылка C' следует из предпосылки C, т. е. $C' \subset C$. Заметим, что композиционный вес $W(C_i^* | \, C, \mathcal{R})$ фактически представляет собой оценку условной вероятности $\mathbf{P}(C_i^* | \, C)$. Тогда множество правил $\mathcal{R}$ индуцирует некоторый классификатор $f_{\mathcal{R}} : \mathrm{X} \to D$ такой, что для $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathrm{X}$

$$f_{\mathcal{R}}(\mathbf{x}) = \arg \max_i W(C_i^* | \, C(\mathbf{x}), \mathcal{R}),$$

где $C(\mathbf{x}) = \{X_1 = x_1\} \cap \{X_2 = x_2\} \cap \ldots \cap \{X_n = x_n\}$. Пусть теперь $\mathbf{K}$ – некоторый статистический критерий проверки нулевой гипотезы $H_0 : \mathbf{P}(C_i^* | \, C) = \theta_0$ против двусторонней альтернативы $H_1 : \mathbf{P}(C_i^* | \, C) \neq \theta_0$, $\Delta = \{\mathbf{P}(C_i^* | \, C),\ i \in D\}$ – некоторое семейство *допустимых* условных вероятностей, а система правил $\mathcal{R}$ такова, что для любой допустимой условной вероятности $\mathbf{P}(C_i^* | \, C)$ статистический критерий $\mathbf{K}$ не отвергает нулевую гипотезу

$$H_0 : \mathbf{P}(C_i^* | \, C) = W(C_i^* | \, C, \mathcal{R})$$

против соответствующей двусторонней альтернативы. Такие системы правил, фактически представляющие собой вероятностно-статистическую модель семейства допустимых условных вероятностей Δ, были использованы для построения классификаторов. Для их индуктивного построения применялась система «Конструктор правил» [9].

Результаты экспериментов. С целью проверки работоспособности описанного выше подхода для определения авторства литературных текстов был проведен ряд

экспериментов. В качестве исходного материала для формирования обучающих и тестовых выборок (необходимых для индуктивного построения классификаторов и оценки величины p_0) были использованы публицистические статьи Ф. М. Достоевского, опубликованные в журналах «Время» и «Эпоха» в период с 1861 по 1865 г. Количество статей составило 11, объем каждой из них – от 6 до 668 предложений и общий объем – 1819 предложений. В качестве альтернативных выступали авторы, сотрудничающие в это время в указанных журналах (М. М. Достоевский, А. А. Григорьев и др.). Количество их статей составило 11, общий объем – 1121 предложение.

Методика проведения каждого эксперимента заключалась в следующем:

1) в качестве контрольной выбиралась некоторая статья из совокупности статей Ф. М. Достоевского и альтернативных авторов;

2) оставшаяся совокупность статей делилась на обучающую и тестовую выборки, причем тестовая состояла только из предложений, принадлежащих Ф. М. Достоевскому;

3) по обучающей выборке осуществлялось индуктивное построение системы правил $\mathcal{R}$ (и соответствующего классификатора $f_{\mathcal{R}}$) рассмотренного выше вида;

4) по тестовой выборке с использованием построенного классификатора производилась оценка величины p_0;

5) для контрольной статьи вычислялись статистика критерия $N_{f_{\mathcal{R}}}$, соответствующее p-значение и осуществлялась проверка нулевой гипотезы, что автором контрольной статьи является Ф. М. Достоевский.

Поскольку общее число статей было 22 и в каждом эксперименте одна статья являлась контрольной, то общее количество экспериментов, проведенных по вышеуказанной методике, также составило 22. Результаты экспериментов следующие. Из 11 статей Ф. М. Достоевского гипотеза о его авторстве была отвергнута для двух статей на 5%-ном уровне значимости и для одной на 1%-ном уровне. Для контрольных статей альтернативных авторов гипотеза об авторстве Ф. М. Достоевского была отвергнута для всех 11 статей на 1%-ном уровне значимости. Таким образом, в 22 экспериментах было принято два неверных решения при уровне значимости 5% и одно при уровне значимости 1%. Учитывая небольшие (со статистической точки зрения) объемы обучающих, тестовых и контрольных выборок результат, по нашему мнению, можно считать хорошим. Заметим, что оценки p_0, представляющие собой оценки вероятностей неправильной классификации предложений Ф. М. Достоевского (с использованием классификатора $f_{\mathcal{R}}$) варьировались в различных экспериментах от 0.38 до 0.41.

Описанный подход был использован для проверки гипотез об авторстве некоторых литературных работ с неустановленным авторством. Были рассмотрены 22 публицистические статьи и заметки, опубликованные в журналах «Время» и «Эпоха» с 1861 по 1865 г., объемом от 7 до 311 предложений. Для каждой статьи проверялась гипотеза, что она была написана Ф. М. Достоевским. Результаты исследования были переданы специалистам по литературному творчеству Ф. М. Достоевского.

Заключение. По нашему мнению, представленный подход имеет некоторые преимущества по сравнению с традиционными методами. Основным его преимуществом является синтез всех доступных лингвостатистических параметров в одном критерии. Проведенные эксперименты показали, что данный подход работоспособен даже в случае коротких литературных текстов, когда применение других методик малооправдано.

Литература

1. *Фукс В.* По всем правилам искусства // Искусство и ЭВМ / пер. с нем.; под ред. Ф. Я. Фридмана. М.: Мир, 1975. 557 с.

2. *Хетсо Г.* Принадлежность Достоевскому: К вопросу об атрибуции Ф. М. Достоевскому анонимных статей в журналах Время и Эпоха. Oslo: Solum Forlag A. S., 1986. 82 с.

3. *Бороджин Л. И., Милов Л. В., Морозова Л. Е.* К вопросу о формальном анализе авторских особенностей стиля в произведениях Древней Руси // Математические методы в историко-экономических и историко-культурных исследованиях: сб. статей / под ред. Н. Д. Ковальченко. М.: Наука, 1977. С. 298–326.

4. *Синилева А. В.* Атрибуция «Романа с кокаином»: лингвостатистическое исследование. Нижний Новгород: Изд-во Нижегородск. гос. ун-та им. Н. И. Лобачевского, 2000. 92 с.

5. *Хмелев Д. В.* Распознавание автора текста с использованием цепей А. А. Маркова // Вестн. Моск. ун-та. Сер. 9: Филология. 2000. № 2. С. 115–126.

6. *Марусенко М. А., Бессонов Б. Л., Богданова Л. М.* и др. В поисках потерянного автора: Этюды атрибуции / под ред. М. А. Марусенко. СПб.: Филологич. ф-т С.-Петерб. ун-та, 2001. 216 с.

7. *Рогов А. А., Сидоров Ю. В., Солопова А. И., Суровцова Т. Г.* Информационно-аналитическая система «СМАЛТ» // Компьютерная лингвистика и интеллектуальные технологии: Труды междунар. конференции «Диалог-2007». М., 2007. С. 470–474.

8. *Hajek P.* Combining Functions for Certainty Factors in Consulting Systems // Intern. J. Man–Machine Studies. 1985. Vol. 22. P. 59–76.

9. *Чистяков С. П.* Применение метода структурной минимизации эмпирического риска при индуктивном построении баз знаний // Труды Ин-та прикл. мат. исследований Карельск. науч. центра РАН. 2002. Вып. 3. С. 213–225.

Статья рекомендована к печати проф. Л. А. Петросяном.

Статья принята к печати 5 марта 2009 г.