УДК 519.68

В. М. БОНДАРЕВ, В. И. РУБЛИНЕЦКИЙ, В. Л. СИГАЛОВ

ПРОГРАММА АВТОМАТИЧЕСКОЙ КЛАССИФИКАЦИИ ТЕКСТОВ

Постановка задачи. Автоматическая классификация текстов по темам — полезная прикладная задача инженерной лингвистики. Эта задача достаточно хорошо изучена [1—3] и обычно решается на основе распределения ключевых слоев, подбираемых разными авторами по разным критериям. Описываемая ниже программа автоматической классификации текстов (ПАКТ) разработана в целом в рамках известных подходов. Однако известные подходы были модифицированы и дополнены с целью обеспечить независимость от языка (допускаются тексты на одном, но любом языке при условиях, что возможно разделить тексты и слова в них, причем тексты записаны в алфа-

витно-числовых символах — для определенности в символах кода ДКОИ); способность к обучению на сравнительно малых наборах текстов с возможностью дальнейшего дообучения; способность классифицировать тексты малого объема (50—150) слов); способность классифицировать тексты быстро (не медленнее, чем они передаются по телетайпу) даже на медленных машинах, типа ЕС-1020.

Задача ставится следующим образом. Пусть дано множество текстов $\{\tau_i\}$; человек-классификатор разделяет их по смыслу на темы — подмножества T_i, $i \in I = \{1, 2, ..., t\}$. Подмножества являются типичным примером нечетких множеств [4]. Автоматический классификатор (АК) может разделить множество $\{\tau\}$ на обычные «четкие» множества; естественно, нужно считать, что АК работает хорошо, если результат его классификации — множества S_i — хорошо совпадает с T_i. Критерий совпадения, однако, нечеток, ибо нечетки множества T_i — разделение на них, по крайней мере, в граничных точках, субъективно и неустойчиво даже для одного субъекта. Чтобы оценить работу АК, был использован следующий прием: выбирался, по возможности, компетентный человек-классификатор (ЧК) и ему предлагалось дать единственное, четкое разбиение $\{\tau\}$ на классы T_i; затем АК получал формальное разбиение S_i, после чего ЧК предъявлялись все тексты, которые АК классифицировал иначе, чем ЧК. При предъявлении конкретного текста реакция ЧК была двух видов; либо «Да, так тоже можно», либо «Нет, это неверно». Решение АК, вызвавшее первую реакцию, мы назовем расхождением, а вторую — ошибкой. При разработке ПАКТ мы стремились минимизировать число ошибок.

Обсуждение метода. Эффективность метода зависит от критерия отбора ключевых слов, от выбора объема и структуры словаря V ключевых слов, от вида решающего правила классификации.

Рассмотрим сначала критерий отбора слов в словарь V. Ключевые слова должны быть достаточно частыми и достаточно хорошо отличать темы друг от друга. К сожалению, частых слов, которые бы отличали данную тему от всех остальных, почти не бывает. Более реалистично искать для каждой пары тем такие ключевые слова, чтобы они отличали одну тему пары от другой. Если потом окажется, что ключевое слово отличает несколько пар, то тем лучше — словарь V будет короче.

Чтобы вывести критерий отбора, введем обозначения: V — число ключевых слов в V; k — номер ключевого слова в V, $k \in K\{1, 2, ..., v\}$; n_i — число слов во всех текстах темы i (короче, в теме i), слово считается столько раз, сколько оно встретилось; n — общее число слов в текстах обучения; π_i — оценка априорной вероятности темы, $\pi_i = n_i/n$ (1); d_k — число раз, которое k-е ключевое слово встретилось во всем корпусе обучения; $d(k/i)$ — число раз, которе k-е ключевое слово

встретилось в теме i; $p(k/i)$ — оценка условной вероятности встретить k-е ключевое слово в теме i.

$$p(k/i) = \frac{d(k/i)}{n_i}, \tag{2}$$

$p(i/k)$ — оценка условной вероятности того, что текст относится к теме i, если в нем встретилось k-е ключевое слово.

Последнюю из введенных величин получим по формуле Байеса. Именно для любой темы

$$p(i/k) = \frac{\pi_j p(k/j)}{\sum_i \pi_i p(k/i)} = \frac{d(k/j)}{\sum_i d(k/i)} = \frac{d(k/j)}{d_k}. \tag{3}$$

Второе равенство получено после подстановки π_i и $p(k/i)$ по формулам (1), (2) и сокращений.

Пусть в некотором тексте τ k-е ключевое слово встретилось r_k раз ($r_k \geqslant 0$). Будем считать, что вероятность встретить в τ k-е ключевое слово не зависит от наличия в τ других ключевых слов или других экземпляров k-го ключевого слова. (Это допущение, особенно во второй части, не вполне корректно, но заменить его нечем; в конечном счете правильность допущения будет подтверждена общей эффективностью классификатора). Тогда вероятность того, что $\tau \in T_i$, равна

$$\Phi(i) = \prod_k p(i/k)^{r_k} = \prod_k \frac{d(k/i)^{r_k}}{d_k^{r_k}}.$$

Переходя для удобства счета к логарифмам, имеем

$$F(i) \equiv \lg \Phi(i) = \sum_k r_k (\lg d(k/i) - \lg d_k). \tag{4}$$

По принципу максимального правдоподобия будем относить текст к той теме, для которой функция $F(i)$ максимальна.

Пусть имеются две темы i и j, $i, j \in I$, а также ψ слов-кандидатов на ключевое слово для их различения. Слово w из множества $W = \{w_1, w_2, \ldots, w_\psi\}$ дает вклад в (4):

$$r_w (\lg d(w/l) - \lg d_w), \quad l = i, j$$

и различает эти темы на величину

$$\Delta_w(i, j) = r_w (\lg d(w/i) - \lg (w/j)). \tag{5}$$

Поэтому предпочтительнее включить в словарь V такое слово w, чтобы

$$\hat{\Delta} \equiv |\Delta_w(i, j)| = \max_{w \in W} \left(r_w \left| \lg \frac{d(w/i)}{d(w/j(} \right| \right).$$

Теперь остается оценить r_W. Возможность классификации текстов основана на том, что статистика распределения ключевых

слов в отдельном тексте $\tau(\tau \in T_i)$ такая же, что и во всей теме i. Поэтому

$$r_w \sim \frac{d(w/i) + d(w/j)}{n_i + n_j}.$$

Поскольку при отыскании $\overset{\cap}{\Delta}$ знаменатель не существен, получаем следующее правило отбора ключевых слов в $V(i, j)$ для той части словаря V, которая заведует различением тем i и j: слова в V отбираются в порядке убывания критерия

$$\gamma = (d(w/i) + d(w/j) \cdot \left| \lg \frac{d(w/i)}{d(w/j)} \right|. \tag{6}$$

Числитель и знаменатель подлогарифмического выражения могут быть равны нулю. В этом случае к числителю и знаменателю добавляется по 0,5. Критерий γ хорошо зарекомендовал себу при классификации, когда $n_i \approx n_j$. При $n_i \ll n_j$ или $n_i \gg n_j$ наблюдалась тенденция отбора в V слов, характерных для «богатой» темы. Чтобы уравновесить ситуацию, мы внесли в предлогарифмический множитель предлогарифмическую поправку и получили модифицированный критерий

$$\gamma' = \left(\frac{d(w/i)}{n_i} + \frac{d(w/j)}{u_j} \right) \lg \left| \frac{d(w/i)}{d(w/j)} \right|. \tag{7}$$

Отметим, что размер словаря V не должен быть велик по двум причинам. Первая вытекает из требования быстрой классификации текстов, а вторая состоит в том, что в большой словарь неизбежно попадут достаточно редкие слова и при малом корпусе слов обучения у них будут недостоверные частоты. Опыт применения ПАКТ показал, что удачная длина парного словаря $|V(i, j)| = 50$. Полный размер словаря V получается существенно меньше, чем $50\,C_t^2$, поскольку многие слова входят в несколько парных словарей. Структура словаря V следующая: это — лексикографически упорядоченный список из V слов, составляющих теоретико-множественное объединение всех парных словарей $V(i, j)$; символьной части соответствует числовая — массив из V строк и C_t^2 столбцов, где в k-й строке и (i, j)-м столбце стоит число

$$V(k; i, j) = \begin{cases} \lg d(k/i) - \lg d(k/j), & w_k \in V(i, j); \\ 0 \text{ в противном случае.} \end{cases} \tag{8}$$

Что касается решающего правила, то при наличии t тем оно состоит в вычислении некоторых параметров полного орграфа K_t. При рассмотрении подлежащего классификации текста τ сравнением с символьной частью V определяются r_k, а затем — все дуги из i в j:

$$c(i, j) = \sum_{w \in V} \Delta_w(i, j) = \sum_{w \in V(i,j)} \Delta_w(i, j). \tag{9}$$

Содержательно $c(i, j)$ означает, насколько вероятность события $\tau \in T_i$ больше (или меньше, в зависимости от знака), чем событие $\tau \in T_j$. Из (5) видно, что $c(i, j) = -c(j, i)$; второе равенство верно, так как $V(k; i, j) = 0$ при $W \in V(i, j)$. Затем определяются веса вершин

$$e_i = \sum_{w \in I \setminus i} c(i, u). \qquad (10)$$

В заключение этого пункта сделаем одно замечание относительно состава словаря V. Выше мы видели, что влияние на классификацию частоты ключевого слова и его различающей силы $d(w/i)$ не равносильны: первое относится ко второму, как число к своему логарифму. Поэтому очень частые слова (предлоги, союзы и т. п.), которые не различают тем из-за неполноты статистики, имеют несколько различные $d(w/i)$, поэтому они включаются в словарь V и вносят существенную путаницу в классификацию (такие слова Р. Г. Пиотровский [1] назвал антипризнаками). Поэтому нужно заранее составить словарь антипризнаков — в этом состоит единственная настройка на язык — и не допускать их в V.

Структура алгоритмов и характеристики программы. Алгоритм состоит из двух частей: обучения (создания словаря V) и классификации. Обучение некритично ко времени, на него затрачивается около минуты (ЭВМ ЕС-1022, 80000 операций/с) на каждые 160 слов обучающего корпуса; классификация осуществляется практически «мгновенно». При обучении вводятся тексты обучения с «ответами» и формируется лексикон — список всех встречных слов без повторений — с накоплением $d(w/i)$. Затем по критерию γ (или γ') формируется словарь V, причем антипризнаки отсеиваются. Дообучение принципиально не отличается от обучения и состоит в пополнении статистики по $d(w/i)$. В ходе дообучения (реже это может случиться и в первоначальном обучении) может не хватить памяти для новых слов. Тогда новым словам позволено вытеснять заданную долю старых слов с наименьшим значением γ (или γ') хотя старые слова могут быть все же лучше новых.

Классификация ведется с использованием словаря V по формулам (9), (10).

Разумеется, в алгоритме использованы экономные процедуры сортировки, дихотомического поиска слова и т. п. На этих деталях мы останавливаться не будем.

Программа написана на ПЛ/1 и рассчитана на длину лексикона до 3640 слов, число тем $t \leqslant 8$. В этом виде программа требует 160К байтов памяти и стандартный набор устройств ЕС ЭВМ, включая НМД типа ЕС-5061.

Результаты тестирования программы. Программа испытывалась на англоязычных текстах, взятых из периодической печати. Был подготовлен словарь антипризнаков из 200 самых

частых служебных слов. Для первоначального обучения вводилось около 6000 слов (40 текстов), неравномерно распределенных по пяти темам. Для парных словарей задавался размер 50, общий объем словаря V составил 151 слово (а не $500 = 50 \cdot c_s^2$) из-за пересечения парных словарей. Чтобы затруднить классификацию, в число тем включалась тема «путаница», куда специально подбирались тексты, похожие по набору слов на тексты других тем, но отличные от них по смыслу.

На экзамене программе предложили 10 текстов, не входящих в материал обучения, — по два на каждую тему. Классификация с использованием словаря V и V' (отобранными соответственно по критерию γ и γ') дала по одной ошибке.

Дообучение производилось на текстах примерно такого же объема, как и при первоначальном обучении. Словарь V увеличился принципиально на 20 слов. На экзамене с предъявлением 10-ти текстов снова была сделана одна ошибка. Однако при использовании V' не было ни одного расхождения, а при использовании V их — два.

Чтобы проиллюстрировать характер классифицировавшихся текстов и понятие расхождения, приведем пример текста

Actor Ned York, who played small parts in th Starsky and Huten TV series, failed yesterday in a strange bid for notoriety — as a killer. He was arrested as a suspect in the hunt for the Los Angeles strangler who has killed 12 girls. York — 6 ft 4 in tall and 17 stone — got himself into the action by making a weird call to detectives. But after questioning him for hours, police dismissed him as a just another crank with a phoney confession — the fourth since the hunt began last October.

Этот текст ЧК отнес к теме «путаница», а АК — к теме «криминальная хроника».

Список литературы: 1. *Пиотровский Р. Г.* Текст, машина, человек. Л., 1975. 327 с. 2. *Попеску А. Н.* Автоматическое индексирование и аннотирование научно-технических текстов//Тр. Всесоюз. семинара по информ. языкам. Секция семиотики. М., 1972. Вып. 4. С. 19—21. 3. *Перцовая Г. М.* Автоматический анализ содержания текста (на материале текстов по стоматологии). Л., 1975. 195 с. 4. *Орлов А. И.* Задачи оптимизации и нечеткие переменные. М., 1980. 63 с.

Поступила в редколлегию 14.07.81