Анализ кластерных конфигураций в одной проблеме фильтрации спама

Дьяконов А. Г.
djakonov@mail.ru
Москва, ВМиК, МГУ им. М. В. Ломоносова

В докладе представлен метод настройки спам-фильтра, который был разработан для участия в соревновании «ECML/PKDD 2006 Challenge» [1] и занял там четвертое место. Конкурсная задача имела две особенности:

1. Для анализа предоставлена только частотная информация (какое слово сколько раз встречается в письме), нет структурной информации (в какой последовательности слова входят), нет контекстной информации (описание заголовка письма, входные данные и т.д.).
2. Обучение происходит на данных из спамовых ловушек (spam traps), а контроль на электронных письмах, которые приходят реальным пользователям. Таким образом, нет гарантии сходства распределений писем в обучении и контроле.

Постановка задачи

Заданы матрицы $S = \|s_{ij}\|_{N_s \times T}$, $M = \|m_{ij}\|_{N_m \times T}$, $U = \|u_{ij}\|_{N_u \times T}$, N_s — число спамовых писем в ловушке, N_m — число нормальных писем в ловушке, N_u — число всех писем в ящике пользователя (эти значения достигают нескольких тысяч), T — число знакомых слов (достигает нескольких сотен тысяч), $s_{ij} = p$ тогда и только тогда, когда j-е слово входит в i-е спамовое письмо p раз (аналогично, элементы m_{ij} описывают вхождения в нормальные письма, u_{ij} — в контрольные). Задача состоит в построении алгоритма, который классифицирует контрольные письма, описанные в U («спам» или «норма»).

Стандартные алгоритмы показывают достаточно плохое качество классификации, и это не связано с эффектом переобучения [2]. Проблема заключается в выборе хорошего пространства признаков. Например, в простейшем пространстве ⟨длина письма, число различных слов⟩ при переходе к контролю классы «меняются местами»: спамовые письма в ящиках пользователя имеют значения этих признаков такими же, как нормальные письма в спам-ловушках, и наоборот.

Формирование пространства признаков

Пусть, для простоты, письмо (строка матрицы U) содержит слова с идентификаторами $1, \ldots, r$ в количестве $c_1, \ldots, c_r$ (соответственно). Рассмотрим матрицу S (для матрицы M все делается аналогично). Рас-

смотрим преобразование $F(H(G(S)))$, где

$$G(\|s_{ij}\|_{N_s \times r}) = \|g(s_{ij}, s_{i1}, \ldots, s_{iT})\|,$$

$$H\big(\|g_{ij}\|\big) = \big(h(g_{11}, \ldots, g_{1N_s}), \ldots, h(g_{r1}, \ldots, g_{rN_s})\big).$$

Функция G осуществляет построчные преобразования матрицы и ее «обрезание» (оставляет столбцы, соответствующие словам письма). Примеры построчных преобразований: нормировка (деление числа вхождения на число слов в письме), «обезличивание» (замена ненулевых чисел единицами). Функция H «схлопывает» матрицу, получая вектор (например, суммирует элементы по вертикали или находит максимальный элемент в каждом столбце). Функция h должна быть монотонной. Функция F — монотонная функция от строки, например сумма элементов или скалярное произведение на вектор $(c_1, \ldots, c_r)$.

Признаки ищем в виде (где f — значение на рассматриваемом письме)

$$f = c_1 F_1^s(H_1^s(G_1^s(S))) - c_2 F_2^m(H_2^m(G_2^m(M))),\ c_1, c_2 \in \mathbb{R}^+$$

с помощью генетического алгоритма. Определив наиболее удачные классы преобразований F, H, G, часто удается провести полный перебор в этих классах, или даже найти аналитические выражения для параметров.

Качество пространства признаков

Качество пространства $[f_1, \ldots, f_n]$ оценивается следующим образом. На множестве писем в ловушке осуществляется кластеризация каким-то фиксированным методом. Этим же методом осуществляем кластеризацию на контроле (для писем пользователя). Предполагаем, что в хороших признаковых пространствах

1) в кластеры входят объекты только одного класса («почти» одного);
2) кластерные конфигурации ловушки и ящика (ящиков) совмещаются друг с другом несложным преобразованием (параллельный перенос, поворот);
3) это преобразование легко определяется по помеченной кластерной конфигурации ловушки (известна классификация) и непомеченной кластерной конфигурации ящика.

На практике чаще всего требуется устойчивость относительно параллельных переносов [3]. Хотя, например, в некоторых задачах из области BCI [4] при применении аналогичной техники наблюдается поворот кластерной конфигурации относительно центра координат.

Формализация условий 1–3 и является «ядром метода». Для решения задачи предлагается оценивать устойчивость кластерных концигураций, и выбирать пространства, в котором они наиболее устойчивы.

Работа выполнена при поддержке РФФИ, проект № 05-01-00332, Минобрнауки РФ, гранта Президента РФ, МК-533.2007.9.

Литература

[1] `www.ecmlpkdd2006.org/challenge.html`.

[2] *Воронцов К. В.* Обзор современных исследований по проблеме качества обучения алгоритмов // Таврический вестник информатики и математики. — Симф.: 2004. — № 1.— С. 5–24.

[3] *Дьяконов А. Г.* Об одном подходе к решению задач из области BCI // Докл. XII всеросс. конф. ММРО-12, М.: МАКС Пресс, 2005. — С. 95–97.

[4] *Jose del R. Millan* Brain-computer interfaces // M. A. Arbib (ed.), Handbook of Brain Theory and Neural Networks, 2nd ed., Cambridge: MIT Press, 2002.