Математические структуры и моделирование
2009, вып. 20, с. 107–112

УДК 004.492.3

МЕТОД КЛАССИФИКАЦИИ ГИСТОГРАММ ДЛЯ ФИЛЬТРАЦИИ СПАМ-ИЗОБРАЖЕНИЙ

Д.А. Лыфарь, В.В. Коробицын

A method for fast classification of spam images based on using machine learning AdaBoost classifier for color and grayscale image histograms is presented. False positives and false negatives are evaluated. Suggestions about using the method as a filter in real antispam engine are given.

1. Введение

Современные методы фильтрации спама основываются не только на анализе содержимого тела письма, поскольку методы рассылки спама совершенствуются. Даже человек по содержанию письма не сразу способен отличить спам-сообщения от обычных писем. Так, если раньше простейший фильтр Байеса, предложенный в 2002 году Полом Грэмом [1], позволял существенно снизить объем спама с очень низким количеством позитивных промахов, то в настоящее время этого недостаточно. Существует множество способов для спам-сообщений обойти фильтрацию фильтром, работающим на основе анализа текстового содержимого письма. К их числу относятся: рассылка спам-сообщений с графическими вложениями; сообщений, тело которых содержит не рекламный текст, а представляет собой обычное письмо с ссылками на рекламные сайты. В последнее время участились случаи взлома злоумышленниками сайтов, IP адрес которых имеет хорошую репутацию, и помещении страниц на взломанный сайт, в которых содержится перенаправление (redirect) на рекламный. Далее в спам-сообщение вставляется подобная ссылка, которая не проходит ни фильтрацию по содержимому, ни фильтрацию по IP/URL. Выявление подобных атак возможно сводится к постобработке данных письма и обучении антиспам-базы.

В данной статье рассмотрен вид графических спам сообщений и предложен метод их фильтрации.

2. Характерные признаки спам-изображений

Характерным примером графического спама может служить изображение, приведенное на рисунке 1а. Для фильтрации подобных изображений обычно приме-

Омский государственный университет.
E-mail: dlyfar@gmail.com

няется метод оптического распознавания символов. Примером применения подобного метода фильтрации может служить известный сервис почты gmail [2]. Однако оптическое распознавание требует много вычислительных ресурсов и становится неэффективным, когда изображение пропускается через ряд графических фильтров с целью искажения. Так же подобные фильтры обычно натренированы на распознавание ограниченного множества языков [3].

а) б)

Рис. 1. Изображения: а) спам, б) обычное

Предлагаемый метод работает вне зависимости от языка или степени искаженности текста. Стоит заметить, что этот метод фильтрации может быть использован в качестве основной оценки, однако чтобы снизить число изображений, которые распознаны неверно — необходимо подкрепить результат фильтра рядом других признаков, говорящих о принадлежности письма к классу спам-сообщений (например, фильтрация по IP из заголовка письма).

Заметим, что в отличие от нормального изображения в спам-изображении большую часть занимает текст. Если построить grayscale гистограмму для обычного и спам-изображений станет очевидным их различие по распределенности компонент на гистограмме (при построении гистограмм использовалось число корзин B=64). На рисунке 2 заметно, что у типичного спам-изображения присутствует несколько пиков, в то время как у обычного изображения компоненты гистограммы распределены равномерно (см. рис. 3). Это один из признаков, который будет учитываться в принятии решении о классификации данного изображения.

Фильтрация на основании данных grayscale гистограммы особенно эффективна, когда спам-изображение содержит в себе в основном текст. Те спам-изображения, в которых присутствует не только текст, лучше поддаются классификации на основании данных цветовых гистограмм (в RG или HV-пространстве). Мы считаем изображение спамом, когда оба классификатора имеют этот результат, чтобы уменьшить число позитивных промахов. Позитивным промахом (false positive) принято считать ситуацию, в которой классификатор ошибочно считает нормальное письмо спамом, негативным промахом (false negative) принято считать классификацию спама как нормального изображения.

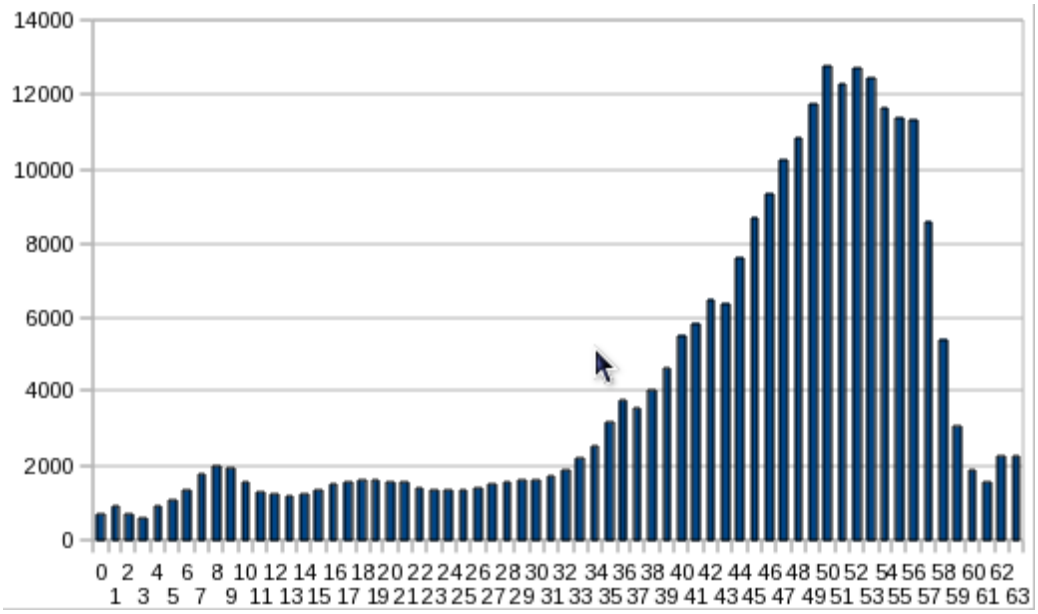

Рис. 2. Grayscale гистограмма для спам-изображения

3. Классификация изображений

В этой работе мы использовали классификатор AdaBoost. Сначала мы должны обучить классификатор уже известным нормальным и спам-изображениям, чтобы сформировать базу данных изображений, на основании которой будут делаться предположения о классе данного изображения. Алгоритм работы AdaBoost подробно описан в [4], здесь мы приводим лишь краткое описание. Этот алгоритм был успешно использован во многих областях, в частности для задачи поиска лиц на изображении.

Требуется построить классифицирующую функцию $F : X \to Y$, где X - пространство векторов признаков (в нашем случае это данные grayscale и цветовых гистограмм), Y - пространство меток классов (в нашем случае это два класса: спам- и нормальное изображение). Пусть в нашем распоряжении имеется обучающая выборка $(x_1, y_1), ..., (x_n, y_n)$, где $x_i \in X$ — вектор признаков, а $y_i \in Y$ — метка класса, к которому принадлежит x_i. Далее в статье мы будем рассматривать задачу с двумя классами, то есть $Y = \{-1; +1\}$. Также у нас есть семейство простых классифицирующий функций $H : X \to Y$. Мы будем строить финальный классификатор в следующей форме:

$$F(x) = \sum_{m=0}^{M} \alpha_m h_m(x). \tag{1}$$

Построим итеративный процесс, где на каждом шаге будем добавлять новое слагаемое

$$f_m = \alpha_m h_m(x), \tag{2}$$

вычисляя его с учётом работы построенной части классификатора (f_0, f_1,..., f_{n-1}). Приведем псевдокод алгоритма AdaBoost:

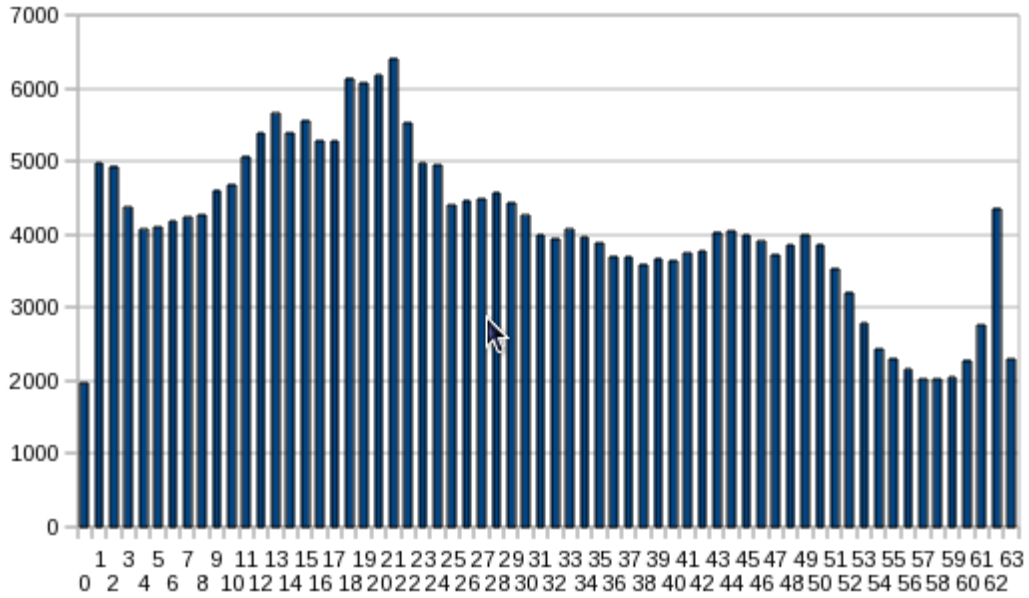

Рис. 3. Grayscale гистограмма для обычного изображения

1. Пусть задана обучающая выборка $(x_1, y_1), \ldots, (x_N, y_N)$ и распределение весов $D_1(i) = 1/N$.

2. Для каждого шага $m = 1, 2, \ldots, M$ выполнить:

   а) выбрать наилучший для текущего распределения $D_m(i)$ слабый классификатор $h_m(x) \in H$ по формуле

   $$h_m = \arg\min_{h_j \in H} \epsilon_j = \sum_{i=1}^{N} D_m(i)[y_i \neq h_j(x_i)];$$

   б) вычислить коэффициент $\alpha_m = \frac{1}{2}\log(\frac{1-\epsilon_m}{\epsilon_m})$;

   в) запомнить $f_m(x) = \alpha_m h_m(x)$ и обновить распределение

   $$D_{m+1}(i) = \frac{D_m(i)\exp(-\alpha_m y_i h_m(x_i))}{Z_m},$$

   где Z_m — нормирующий коэффициент, обеспечивающий выполнение условия $\sum\limits_{i=1}^{N} D_{m+1}(i) = 1$.

3. Составляем итоговый классификатор:

   $$F(x) = \operatorname{sgn}\left[\sum_{m=1}^{M} f_m(x)\right].$$

4. Результаты эксперимента

Мы использовали реализацию алгоритма AdaBoost из открытой библиотеки opencv. В качестве набора для обучения было использовано 946 нормальных изображений и 825 спам-изображений. Полученный классификатор работал на реальном потоке электронных сообщений одного из провайдеров Европы, результаты потока помогли составить приблизительную оценку эффективности этого метода. Как уже было сказано выше, под эффективностью метода фильтрации понимается процент позитивных (FP) и негативных (FN) промахов.

Эксперименты показали, что эффективность предсказания зависит от числа корзин (B=64 — это эмпирически подобранное значение, с дальнейшим ростом B эффективность метода практически не изменялась), от размера изображения (все изображения приводились к размеру 512x512 пикселей). Результаты показали $FP = 0.014$ и $FN = 0.12$ для 100 изображений из потока. При этом были выявлены следующие преимущества и недостатки метода.

Преимуществами можно считать высокую производительность и низкий процент FP. Производительность метода значительно выше, чем у методов, использующих алгоритмы распознавания текста. Достаточно низкий процент FP позволяет использовать этом метод в качестве вспомогательного фильтра и в качестве основного во время спам-атак для писем, содержащих только изображения и пришедших с белых IP-адресов.

К недостаткам можно отнести значительно высокое FN. Фильтр ошибочно определяет некоторый набор изображений как спам: сканированные документы, изображения из новостных рассылок. Фильтр слабо реагирует на те рекламные изображения, где большую площадь занимает изображение, идентичное нормальному.

5. Заключение

Представленный метод фильтрации писем целесообразно использовать в цепочке фильтров антиспам-системы для фильтрации спам-изображений. Так как сегодня спам-атаки, содержащие изображения, время от времени представляют собой достаточно высокий процент писем, данный фильтр является оптимальным выбором с точки зрения производительности и качества. Чтобы уменьшить число позитивных промахов, целесообразно будет использовать фильтр в качестве дополнительного при оптическом распознавании символов в той ситуации, когда фильтр классифицирует сообщение как спам.

Литература

1. Graham, P. A Plan for Spam [Электронный ресурс] / P. Graham. – Режим доступа: http://www.paulgraham.com/spam.html (2.10.2009).
2. Gmail uses Google's innovative technology to keep spam out of your inbox [Электронный ресурс] – Режим доступа: http://www.google.com/mail/help/fightspam/spamexplained.html (3.10.2009).

3. Tesseract-ocr - Project Hosting on Google Code [Электронный ресурс] – Режим доступа: http://code.google.com/p/tesseract-ocr/ (3.10.2009).
4. Sochman, J., Matas, J. AdaBoost [Электронный ресурс] / J. Sochman, J. Matas. – Режим доступа: http://cmp.felk.cvut.cz/ sochmj1/adaboost_talk.pdf (3.10.2009).
5. Херн, Д., Бейкер, М.П. Компьютерная графика и стандарт OpenGL. – М.: Издательский дом «Вильямс», 2005.