АНАЛИЗ МЕТОДОВ ФИЛЬТРАЦИИ СПАМА

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шершень А.С.

Серебряная Л.В. – канд. техн. наук, доцент

На сегодняшний день практически каждый человек пользуется электронной почтой и сталкивается с проблемой огромного количества рекламных рассылок. Несколько десятилетий люди искали наилучший метод для фильтрации спама, и сейчас имеется достаточное количество альтернатив. В данной работе будет проведен обзор наиболее популярных методов и проведен их сравнительный анализ на реальных письмах электронной почты.

Наиболее распространённым способом общения в сети Интернет, на сегодняшний день, является электронная почта. Однако при разработке протокола доставки электронной почты не были учтены никакие методы проверки подлинности личности отправителя. Для отправки сообщения достаточно знать лишь электронный адрес получателя, что существенно облегчает и способствует массовой рассылке коммерческих сообщений именно по электронной почте.

Количество писем с коммерческой рекламой с каждым годом возрастает, и на сегодняшний день по оценкам разных экспертов доля спама в мировом почтовом трафике составляет от 70 до 95 процентов.

Огромное количество рекламных сообщений наносит ощутимый вред получателям. Пользователю приходится тратить много времени на отсеивание спама и поиск нужной почты, часто интернет трафик стоит дорого и получателю приходится платить за ненужную почту. Однако наибольший вред от спама заключается в пользователях, открывающих нежелательные письма, распространяющие вирусы, которые, зачастую, продолжают распространение спама. Так же от массовых коммерческих рассылок страдают провайдеры, которым приходится бесконечно наращивать мощности своих центров обработки данных, в условиях всё время растущего объёма бесполезного трафика.

Становится понятна необходимость в эффективных методах фильтрации спама по содержимому писем. Существует множество методов классификации, варьирующихся по простоте реализации, эффективности, производительности и возможности следования за трендом спам рассылок.

В данной работе будут рассмотрены и протестированы следующие методы:

– самоорганизующиеся карты Кохонена (SOM);

– метод опорных векторов (SVM);

– наивная байесовская классификация (MNB);

– деревья решений (BDT).

Самоорганизующиеся карты Кохонена – это частный случай нейросети, способной обучаться без учителя. Основная функция SOM заключается в идентификации функции в n-мерном пространстве и проецировании этого пространства на двухмерную плоскость в соответствии с подобием [1]. SOM трудно использовать после обучения. Хоть визуально некоторые кластеры и проявляются в выходной карте, вычислительно сложно классифицировать новый вход в сформированный кластер и семантически маркировать его принадлежность. Поэтому для классификации используется метод взвешенного мажоритарного ввода, при котором соседи введенного вектора голосуют за принадлежность к их кластеру, чьих голосов окажется больше, к тому кластеру и будет отнесен новый вектор [2].

Метод опорных векторов давно известен хорошими показателями в фильтрации спама. Основная идея метода - перевод исходных векторов в пространство более высокой размерности и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве [3]. Две параллельных гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Разделяющей гиперплоскостью будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше будет средняя ошибка классификатора.

Наивная байесовская классификация – наиболее часто используемый на сегодняшний день метод, показывающий хорошие результаты при простейшей реализации, основанной на теореме Байеса. Сообщение разбивается на слова, и каждое слово оценивается с точки зрения частоты встречаемости в письмах со спамом. Таким образом, суммируя все частоты, можно получить вероятность принадлежности всего письма к спаму.

Алгоритмы деревьев решений работают сверху вниз. От набора признаков вектора входных значений пытаются подобрать функцию, которая лучше всего делит классы, и выбирают её как корневой узел, а ветви корневого узла отражают возможные значения будущих входных векторов [4].

Для сравнения этих методов был выбран наиболее часто используемый исследователями фильтрации спама корпус писем Enron, содержащий письма более чем со 150 ящиков электронной почты [5].

Далее на рисунках 1 и 2 будет представлена статистика по классификации писем данными методами.

На Рисунке 1 представлено процентное соотношение определения спама алгоритмами самоорганизующихся карт (SOM), деревьев решений (BDT) и опорных векторов(SVM). Видно, что практически на всех наборах входных данных преобладает SOM, за исключением некоторых погрешностей.

	SOM	BDT	SVM
Enron 1	87.31	87.07	87.20
Enron 2	91.74	95.20	97.21
Enron 3	94.82	94.44	94.50
Enron 4	85.87	85.55	85.73
Enron 5	97.87	97.87	97.87
Enron 6	94.43	94.39	94.43

Рисунок 1 – Сравнение методов SOM, BDT, SVM

На Рисунке 2 представлена сравнительная характеристика SOM и наиболее используемого на сегодняшний день метода наивной байесовской фильтрации (MNB):

	HAM		SPAM	
	SOM	MNB	SOM	MNB
Enron 1	99.95	95.25	87.31	96
Enron 2	96.46	97.83	91.74	96.68
Enron 3	100	98.88	94.82	96.64
Enron 4	99.45	99.05	85.87	97.79
Enron 5	100	95.64	97.87	99.69
Enron 6	99.86	96.88	94.43	98.1

Рисунок 2 – Сравнение методов SOM и MNB

Из Рисунка 2 можно сделать вывод о преобладании метода наивной байесовской фильтрации (MNB) над методом самоорганизующихся карт (SOM) в определении спама, также можно сделать вывод о большом количестве ошибок первого рода, когда валидные сообщения (HAM) классифицируются как спам, что может довольно негативно сказаться на опыте конечных пользователей при использовании данного метода. В то же время самоорганизующиеся карты Кохонена показывают сравнительно неплохие показатели фильтрации спама и практически нулевой процент ошибок первого рода.

Из проведенного анализа можно сделать вывод о превосходстве самоорганизующихся карт над всеми рассматриваемыми методами в проценте ошибок первого рода и приблизительно равных показателях в проценте классификации спама с методом наивной байесовской фильтрации. Таким образом, есть все основания для введения в широкое использование метода SOM, игнорируя более сложную реализацию.

Список использованных источников:

1. T. Kohonen. Self-Organizing Maps. 2nd edition. — Springer-Verlag, New York, 1997 — ISBN 3-540-67921-9.
2. H. Drucker, D. Wu, V.N. Vapnik. Support Vector Machines for Spam Categorization. — Transactions on Neural Networks, 1999. — 1048-1054 с.
3. T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. — Proceedings of 10th European Conference on Machine Learning, 1998. — 137-142 с.
4. R.E. Schapire. The Strength of Weak Learnability. — Machine Learning, V (2), 1990. — 197-227 с.
5. Androutsopoulos, G. Paliouras, V. Karkaletsi, G. Sakkis, C.D. Spyropoulos, P. Stamatopoulos. Learning to Filter Spam E-Mail: A Comparison of a Naïve Bayesian and a Memory-Based Approach. — Proceedings of the Workshop Machine Learning and Textual Information Access, 2000. — 1-13 с.

АВТОМАТИЗАЦИЯ ЗАДАЧ УЧАСТНИКОВ УЧЕБНОГО ПРОЦЕССА

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Шульга Е.С.

Сурков К.А. – старший преподаватель

В докладе рассмотрены задачи, возникающие перед преподавателями и студентами в процессе учебной деятельности, проанализированы существующие варианты их решения. По результатам рассмотрения их недостатков предложено создание унифицированной системы для участников учебного процесса.