Интернет-журнал «Науковедение» ISSN 2223-5167 http://naukovedenie.ru/

Том 8, №2 (2016) http://naukovedenie.ru/index.php?p=vol8-2

URL статьи: http://naukovedenie.ru/PDF/134TVN216.pdf

DOI: 10.15862/134TVN216 (http://dx.doi.org/10.15862/134TVN216)

Статья опубликована 16.05.2016.

Ссылка для цитирования этой статьи:
Ларионова А.В., Хорев П.Б. Оценка эффективности метода фильтрации спама на основе искусственной нейронной сети // Интернет-журнал «НАУКОВЕДЕНИЕ» Том 8, №2 (2016) http://naukovedenie.ru/PDF/134TVN216.pdf (доступ свободный). Загл. с экрана. Яз. рус., англ. DOI: 10.15862/134TVN216

УДК 004.81

Ларионова Анна Владимировна[1]
ФГБОУ ВО «Российский государственный социальный университет», Россия, Москва[2]
Аспирант
E-mail: tarelo4ka76@mail.ru
РИНЦ: http://elibrary.ru/author_items.asp?authorid=828879

Хорев Павел Борисович
ФГБОУ ВО «Национальный исследовательский университет «Московский энергетический институт», Россия, Москва
Преподаватель
Кандидат технических наук, доцент
E-mail: pbkh@mail.ru
РИНЦ: http://elibrary.ru/author_items.asp?authorid=620811

Оценка эффективности метода фильтрации спама на основе искусственной нейронной сети

Аннотация. В данной статье рассматривается задача фильтрации спама как задача бинарной классификации сообщений электронной почты на два класса: спам и обычная почта, с использованием метода искусственного интеллекта – искусственной нейронной сети. Задача фильтрации спама является актуальной проблемой, так как технологии создания спама развиваются следом за средствами защиты от спама, что требует переосмысления подходов к задаче фильтрации спама и применения методов и средств искусственного интеллекта. В качестве решения проблемы предлагается использовать подход на основе методов искусственного интеллекта, в частности на основе искусственной нейронной сети. Данный подход требует подготовки обучающей и тестовой выборок сообщений для обучения классификатора, выделения значимых признаков сообщений, настройки параметров модели, оценки точности классификатора. В статье описывается постановка эксперимента и его результаты для оценки точности классификатора, на основании которых делается вывод о целесообразности использования методов искусственного интеллекта и, в частности, искусственной нейронной сети при фильтрации спама.

Ключевые слова: фильтрация спама; искусственные нейронные сети; искусственный интеллект; спам; классификация сообщений; выделение признаков сообщений; персептрон;

[1] LinkedIn: https://ru.linkedin.com/in/anna-larionova-74434689

[2] 140180, Российская Федерация, Московская область, г. Жуковский, ул. Семашко, д. 8, корп. 2, кв. 41

теорема Байеса; машинное обучение; информационная безопасность; оценка эффективности; точность классификации

В современном мире, где реклама является двигателем торговли, с развитием сети Internet и средств общения, проблема нежелательной рекламы и сообщений (спам [6]) требует интеллектуального подхода для ее решения. Современные методы борьбы со спамом, основанные на лингвистических сигнатурах, правилах фильтрации сообщений [8], становятся все менее эффективными, так как требуется увеличение трудозатрат специалистов по защите от спама на поддержание этих сигнатур и правил в актуальном состоянии. Таким образом, современные методы борьбы со спамом требуют постоянного участия человека для эффективного анализа текста, они не способны самостоятельно вырабатывать эти правила, то есть самообучаться. Если рассматривать человека как средство борьбы со спамом, то можно сказать, что он обладает способностью обнаружения признаков спама, основываясь на собственном опыте и предпочтениях, знаниях о добровольных новостных и рекламных подписках, обучаемостью, его работа не сводится к шаблонам и потому более эффективна. Именно поэтому задача создания средства борьбы со спамом сводится к наделению средства борьбы со спамом навыками и качествами, присущими человеку: способность к обучению, система предпочтений и исключений, анализ контекста, система принятия решений.

Предлагаемый автором статьи метод фильтрации спама основан на использовании нейронной сети, выступающей в качестве механизма принятия решений, давая на выходе вероятностную оценку «спамности» всего сообщения. Нейронная сеть построена по принципу многослойного персептрона [7, 10] с одним скрытым слоем. Количество нейронов в скрытом слое было определено по формуле Арнольда – Колмогорова – Хехт-Нильсена [4] и равно двенадцати. Во входном слое нейронной сети содержатся четыре нейрона, что соответствует четырем входным параметрам нейронной сети (вероятность наличия спама в сообщении, связность текста сообщения, смысловая направленность текста сообщения, удельное число «обманных» замен символов одного алфавита символами другого алфавита), выделенных при обработке текста входящего сообщения. В выходном слое нейронной сети имеется всего один нейрон, что связано с возложенной на него задачей принятия решения о наличии спама в сообщении.

Оценка эффективности разработанного приложения [5, 9] определит, является ли конкретный рассмотренный подход и способ реализации подхода эффективным, то есть, следует ли использовать его в реальной практике. Так же, оценка эффективности позволит определить целесообразность использования нейронных сетей в задачах защиты компьютерных систем от спама.

Для проведения оценки эффективности разработанное программное средство оценивалось с точки зрения ложных срабатываний и успешных обнаружений спама (ошибки 1-го и 2-го родов). Оценка эффективности проводилась экспериментальным путем. В качестве спам-сообщений эксперимент содержал сообщения, уже помеченные спам-фильтрами как нежелательные, всего для проведения эксперимента потребовалось собрать два типа входных данных: сообщения, текст которых содержит спам, и сообщения, текст которых не является спамом. Далее наборы данных обрабатывались разработанным программным модулем и собиралась статистика ответов. Полученная статистика сравнивалась с ожидаемой, на основании сравнения которых определялись ошибки первого и второго рода – ложные срабатывания и принятие спам-сообщений за обычные. Для оценки эффективности вычислялась относительная величина числа ошибок первого (α) и второго рода (β) к размеру входных данных. Таким образом, сумма относительных величин ошибок первого и второго

рода и относительная величина верно распознанных сообщений (γ) образуют, согласно теории вероятности, полную группу событий, откуда эффективность выражается как γ=1-(α+β).

Репрезентативной выборкой входных данных будем считать выборку, в которой длина текста сообщения составляет не менее десяти и не более тысячи слов, в выборку не входят сообщения на иностранных языках, причем, число обычных сообщений и спам-сообщений в выборке одинаково (при случайном выборе сообщения из представленного набора данных вероятность того, что оно является или не является спамом равна 0,5, то есть события равновероятны). К обычным сообщениям отнесены письма, добровольные новостные и рекламные рассылки, выдержки из научных и научно-популярных статей и художественных литературных произведений.

Результаты эксперимента отражены в таблице 1 «Результаты обработки спам-сообщений в ходе эксперимента».

Таблица 1

Результаты обработки спам-сообщений в ходе эксперимента

№	Ответ программы	Оценка	Спамность	Связность	Направленность	Замены
1	Верный	0,62	0,97	0,77	0,8	0,5
2	Верный	0,55	0,86	0,81	0,8	0,25
3	Верный	0,63	0,81	0,52	0,8	0,5
4	Верный	0,65	0,83	0,77	0,8	0,85
5	Верный	0,52	0,39	0,71	0,8	0,6
6	Верный	0,51	0,26	0,44	0,8	0,5
7	Верный	0,5	0,6	0,8	0,8	0
8	Неверный	0,25	0,08	0,77	0,8	0
9	Верный	0,63	0,9	0,5	0,8	0,43
10	Неверный	0,41	0,61	0,76	0,8	0
11	Верный	0,59	0,5	0,77	0,8	0,81
12	Верный	0,54	0,88	0,67	0,8	0,1
13	Верный	0,65	0,63	0,44	0,8	0,83
14	Верный	0,63	0,77	0,26	0,8	0,33
15	Верный	0,54	0,01	0,5	0,8	1
16	Неверный	0,48	0,5	0,57	0,8	0,2
17	Верный	0,63	0,9	0,7	0,8	0,6
18	Неверный	0,49	0,89	0,82	0,8	0
19	Верный	0,65	0,61	0,61	0,8	0,95
20	Верный	0,67	0,9	0,8	0,8	1
21	Верный	0,63	0,7	0,54	0,8	0,69
22	Верный	0,51	0,61	0,38	0,8	0
23	Верный	0,63	0,98	0,64	0,8	0,5
24	Верный	0,59	0,96	0,38	0,8	0
25	Верный	0,62	0,73	0,51	0,8	0,5
26	Верный	0,52	0,81	0,47	0,7	0
27	Верный	0,65	0,67	0,76	0,8	1
28	Верный	0,55	0,85	0,56	0,8	0,07
29	Верный	0,6	0,75	0,81	0,8	0,66

№	Ответ программы	Оценка	Спамность	Связность	Направленность	Замены
30	Верный	0,59	0,94	0,68	0,8	0,25
31	Верный	0,51	0,92	0,77	0,8	0
32	Верный	0,52	0,93	0,75	0,8	0
33	Верный	0,53	0,7	0,38	0,8	0
34	Верный	0,64	0,68	0,53	0,8	0,75
35	Верный	0,66	0,71	0,69	0,8	1
36	Верный	0,67	0,99	0,79	0,8	1
37	Неверный	0,49	0,89	0,8	0,8	0
38	Верный	0,66	0,89	0,58	0,8	0,82
39	Верный	0,67	0,87	0,68	0,8	1
40	Верный	0,64	0,74	0,49	0,8	0,61
41	Верный	0,67	0,88	0,48	0,8	0,8
42	Верный	0,62	0,74	0,71	0,8	0,66
43	Верный	0,68	1	0,34	0,69	0,8
44	Верный	0,59	0,91	0,86	0,8	0,42
45	Верный	0,62	0,65	0,55	0,8	0,66
46	Верный	0,62	0,65	0,55	0,8	0,64
47	Верный	0,67	0,81	0,67	0,8	1
48	Верный	0,64	0,88	0,8	0,8	0,73
49	Верный	0,67	0,83	0,6	0,8	1
50	Верный	0,62	0,8	0,84	0,8	0,75

Слова «верно» и «неверно», используемые в таблице, означают, обнаружен ли спам в сообщении или же спам-сообщение было пропущено. Поля «Оценка» (решение нейронной сети, принятое на основе входных параметров), «Спамность» (вероятность принадлежности сообщения к спаму, вычисляемая по теореме Байеса [8, 9]), «Связность» (связь частей речи в предложении, определяемая исходя из формализма Бекуса – Наура – упрощенного описания синтаксических конструкций предложения), «Направленность» (определяется по принципу «победитель забирает все», т.е. определяется смысловая категория, к которой принадлежит большинство слов сообщения), «Замены» (количество замен символов одного алфавита символами другого алфавита) относятся к разработанному программному модулю, причем, поле «Оценка» содержит значения, получаемые на выходе нейронной сети (если значение больше ½, то сообщение считается спамом). Значения других полей отображают оценку сообщения по четырем признакам и являются входными параметрами нейронной сети.

В ходе проведения эксперимента на выборке из 50 спам-сообщений, помеченных как «нежелательные» разработанный программный продукт пропустил 5 сообщений. Далее была произведена обработка обычных сообщений, не помеченных как спам. В ходе второго эксперимента при обработке обычных сообщений разработанный программный модуль ошибочно принял за спам 8 сообщений из 50.

Как говорилось выше, эффективность есть количество правильно распознанных сообщений ко всем сообщениям (всего 100 сообщений). Итого общее число обработанных сообщений – 100 штук. Из них ошибочно определено 13. По итогам эксперимента представим среднее значение показателей ошибок первого и второго рода и эффективности разработанного программного модуля 5% ошибок 1-го рода, 8% ошибок 2-го рода, точность классификатора составляет 87%.

Как показал результат проведенного эксперимента, использование нейронных сетей является перспективным направлением в задачах фильтрации спама. Благодаря их способности к обучению и обобщению они способны обнаруживать спам, который не был ранее представлен в обучающей выборке. В отличие от метода сигнатурной фильтрации спама для обнаружения новых видов спама необходимо написание новых сигнатур для его обнаружения.

Так как разработанный программный модуль помимо использования нейронной сети для принятия решения о наличии спама в сообщении использует комплексный подход для определения характеристик сообщения, что позволяет избежать недостатков методов, использующих не более одного-двух признаков спама без учета их зависимости между собой.

В отличие от других методов обнаружения спама, предложенный метод позволяет не только защититься от спама за счет сведений о ненадежных источниках, признаках массовости, а от спама вообще, в том числе, если прислан от доверенных пользователей. Также стоит отметить низкие трудозатраты на поддержание актуальности баз по сравнению с методом обнаружения спама на основе сигнатур. В тоже время в отличие от метода фильтрации на основе теоремы Байеса, нейронная сеть, представленная в предлагаемом методе фильтрации спама, не требует постоянного дообучения – обладает способностью к самообучению [1, 2, 3].

Повышения эффективности разработанного программного модуля без внесения изменений в код программы можно достичь следующими способами:

- формирование новой обучающей выборки с большей размерностью обучающих пар (примеров);
- увеличение словаря.

Разработанный программный модуль имеет следующие перспективы развития:

- создание надстройки для фильтрации спама для почтовых клиентов на основе разработанного программного модуля;
- создание экспертной системы лингвистического анализа, включающей в себя разработанный программный модуль;
- расширение набора поддерживаемых языков;
- добавление возможности пользовательской пометки сообщения как спам вручную для пропущенных разработанным программным модулем спам-сообщений, таким образом, оказывая корректирующее воздействие на нейронную сеть и байесов фильтр в составе разработанного программного модуля;
- применение байесовой фильтрации к словосочетаниям из двух и более слов;
- добавление дополнительного функционала в лингвистический анализ сообщения на основе сигнатур и их автоматической генерации на основе решения нейронной сети для входящего сообщения.

Описанные выше меры позволят по оценкам автора статьи увеличить итоговую эффективность предложенного метода фильтрации спама на основе нейронной сети до 92-96%, что позволит создать на основе разработанного программного модуля конкурентоспособный коммерческий продукт.

ЛИТЕРАТУРА

1. Осовский Станислав. Нейронные сети для обработки информации = Sieci neuronowe do przetwarzania informacji (польск.) / Перевод И.Д. Рудинского. - М.: Финансы и статистика, 2004. - 344 с. - ISBN 5-279-02567-4.
2. Савельев А.В. На пути к общей теории нейросетей. К вопросу о сложности // Нейрокомпьютеры: разработка, применение. - 2006. - №4-5. - С. 4-14. Режим доступа http://www.radiotec.ru/catalog.php?cat=jr7 (открытый).
3. Хайкин С. Нейронные сети: полный курс = Neural Networks: A Comprehensive Foundation. 2-е изд. - М.: Вильямс, 2006. - 1104 с.
4. Ясницкий Л.Н. Введение в искусственный интеллект. М.: Издательский центр «Академия», 3-е издание, 2010 – 176 с.
5. Demsar Janez. Statistical Comparisons of Classifiers over Multiple Data Sets –- 2006. Access link: http://sci2s.ugr.es/sicidm/pdf/2006-Demsar-JMLR.pdf.
6. Mueller Scott Hazen. What is spam? Information about spam. Abuse.net. Retrieved 2007-01-05. Access link: http://spam.abuse.net/overview/whatisspam.shtml (open access).
7. Rosenblatt, Frank. Principles of Neurodynamic: Perceptrons and the Theory of Brain Mechanisms. - М.: Мир, 1965. - 480 с.
8. Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras. Spam Filtering with Naive Bayes - Which Naive Bayes? // Third Conference on Email and Anti-Spam (CEAS). 2006 – 9 p.
9. Vapnik Vladimir N. The Nature of Statistical Learning Theory – 1999. Access link http://web.mit.edu/6.962/www/www_spring_2001/emin/slt.pdf (open access).
10. Warren S. McCulloch, Walter H. Pits. A logical calculus of the ideas immanent in nervous activity. Access link http://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf (open access).

Larionova Anna Vladimirovna
Russian State Social University, Russia, Moscow
E-mail: tarelo4ka76@mail.ru

Khorev Pavel Borisovich
National research institute «Moscow Power Engineering Institute», Russia, Moscow
E-mail: pbkh@yandex.ru

Efficiency evaluating of spam filtering method based on artificial neural network

Abstract. This article discusses the problem of spam filtering as a binary classification task e-mail messages into two classes: spam and regular e-mail, using the method of artificial intelligence - artificial neural network. Spam filtering task is an actual problem, since the technology of spam are advancing together with spam, which requires a rethinking of approaches to the problem of spam filtering and application of artificial intelligence methods. As a solution is proposed to use techniques based on artificial intelligence approach, in particular based on an artificial neural network. This approach requires the preparation of the training and test samples of messages for training the classifier, extraction important features of messages, setting parameters of the model, evaluate the accuracy of the classifier. The article describes an experiment and results of classifier accuracy evaluation, which concludes that the methods of artificial intelligence and, in particular, the artificial neural network with spam filtering feasibility of using are advisible.

Keywords: spam filtering; artificial neural network; artificial intelligence; spam; message classification; feature extraction of messages; perceptron; Bayes theorem; machine learning; information security; efficiency evaluating; classification accuracy

REFERENCES

1. Osovskiy Stanislav. Neyronnye seti dlya obrabotki informatsii = Sieci neuronowe do przetwarzania informacji (pol'sk.) / Perevod I.D. Rudinskogo. - M.: Finansy i statistika, 2004. - 344 s. - ISBN 5-279-02567-4.
2. Savel'ev A.V. Na puti k obshchey teorii neyrosetey. K voprosu o slozhnosti // Neyrokomp'yutery: razrabotka, primenenie. - 2006. - №4-5. - S. 4-14. Rezhim dostupa http://www.radiotec.ru/catalog.php?cat=jr7 (otkrytyy).
3. Khaykin S. Neyronnye seti: polnyy kurs = Neural Networks: A Comprehensive Foundation. 2-e izd. - M.: Vil'yams, 2006. - 1104 s.
4. Yasnitskiy L.N. Vvedenie v iskusstvennyy intellekt. M.: Izdatel'skiy tsentr «Akademiya», 3-e izdanie, 2010 – 176 s.
5. Demsar Janez. Statistical Comparisons of Classifiers over Multiple Data Sets –- 2006. Access link: http://sci2s.ugr.es/sicidm/pdf/2006-Demsar-JMLR.pdf.
6. Mueller Scott Hazen. What is spam? Information about spam. Abuse.net. Retrieved 2007-01-05. Access link: http://spam.abuse.net/overview/whatisspam.shtml (open access).
7. Rosenblatt, Frank. Principles of Neurodynamic: Perceptrons and the Theory of Brain Mechanisms. - M.: Mir, 1965. - 480 s.
8. Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras. Spam Filtering with Naive Bayes - Which Naive Bayes? // Third Conference on Email and Anti-Spam (CEAS). 2006 – 9 p.
9. Vapnik Vladimir N. The Nature of Statistical Learning Theory – 1999. Access link http://web.mit.edu/6.962/www/www_spring_2001/emin/slt.pdf (open access).
10. Warren S. McCulloch, Walter H. Pits. A logical calculus of the ideas immanent in nervous activity. Access link http://www.cse.chalmers.se/~coquand/AUTOMATA/mcp.pdf (open access).