ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 *2006. Nr. 5(69)*
ELEKTRONIKA IR ELEKTROTECHNIKA

TELECOMMUNICATIONS ENGINEERING
T 170
TELEKOMUNIKACIJŲ INŽINERIJA

An Artificial Neural Nets for Spam e-mail Recognition

D. Puniškis
Department of Electronics engineering, Kaunas University of Technology,
Studentu str. 50, LT-51368 Kaunas, Lithuania, phone: +370 686 19904; e-mail: danius.puniskis@stud.ktu.lt
R. Laurutis
JSC „Information Avenue“,
Dvaro str. 55, LT-76344 Siauliai, Lithuania, phone: +370 685 28295; e-mail: remigijusl@aleja.lt
R. Dirmeikis
JSC „TV & Communication Systems“,
Dvaro str. 140, LT-76199 Siauliai, Lithuania, phone: +370 614 73737; e-mail: ramunas@tvc.lt

Introduction

The volume of unsolicited commercial e-mail messages (also called as junk e-mail or spam) transmitted by the Internet has reached epidemic proportions nowadays.

According to MessageLabs statistics [1], there was only 8% of spam of network e-mail traffic in 2001, however it reaches 70% in 2005 and increases exponentially showing high insecurity of internet technologies. An information security incidents met by Lithuanian Internet users are frequent phenomenon too according to the questioning made in portal www.delfi.lt. Even 78% of Internet service users reported had problems with computer virus. The 63% reported coming across spam messages [2]. The computer virus and spam phenomena are the largest disasters of communications industry giving huge annual loss in profit. They also have big potential to penetrate in mobile GSM and UMTS networks. In the late of 2005 started mobile virus era. There have been registered already 15 different type viruses exploiting mobile operating system vulnerabilities in 2005, not including variations of them.

Computer viruses and spaming are closely related. Viruses are developed and used as a tool for collecting as many as possible e-mail addresses, which are later used as correspondence recipients, or even for more evil task. In this way, infected workstations are used as mail proxy servers for spam distribution.

The spam messages raise a lot of problems for internet service providers and users also. Firstly, junk email occupies server storage space and consumes network bandwidth, for second, users are pushed to waste non-trivial amount of time for identifying and removing spam from own computers.

The best solution for avoiding such discomfort would be to develop and refine automatic classifies that can distinguish legitimate e-mail from spam accurately and efficiently.

Spam recognition techniques

For that challenge of technology, many commercial and open-source products exist to accommodate the growing need for spam classifiers, and a variety of techniques have been developed and applied toward the problem, both at the network and user levels.

The simplest and most common approaches are to use filters that screen messages based upon the presence of common words or phrases common to junk e-mail.

Other simplistic approaches include *blacklisting* (automatic rejection of messages received from the addresses of known spammers) and *whitelisting* (automatic acceptance of message received from known and trusted correspondents).

In practice, effective spam filtering uses a combination of these three techniques. The primary flaw of the first two approaches is that they rely on spammers by assuming that they will not change their identities or alter the style and vocabulary of their sales pitches. Whitelisting risks the possibility that the recipient will miss legitimate e-mail from a known or expected correspondent with a heretofore-unknown address, such as correspondence from a long-lost friend.

A variety of text classifiers have been investigated that categorize documents topically or thematically, including probabilistic, decision tree, rule-based, example-based, linear discriminant analysis, regression, support vector machine, and neural network approaches.

However, the problem still exists and there are two main reasons of it. First, the effectiveness of any given anti-spam technique can be seriously compromised by the public revelation of the technique since spammers are aggressive and adaptable. Second, recent variations of

Naïve Bayesian classifiers have demonstrated high degrees of success. In general, these classifiers identify attributes (usually keywords or phrases common to spam) that are assigned probabilities by the classifier [3,4,5].

Dataset analysis

For making up the data set, the first step was to classify manually corpus of 2788 legitimate and 1812 spam emails received over a period of several months. The next step was primary text analysis of spam emails using MS Word text analysis tools. It enabled to find out what kind of spam exploits and techniques are most frequently used by spamers to overcome existing text classifiers [7].

Table 1. The Spam exploit techniques.

Exploit	Spam	Description
Word Obscuring	20%	Misspelling words, putting words into images, etc
Text Chaff	56%	Random strings of characters, random series of words, or unrelated sentences.
Character Encoding	10%	Pharmacy renders into Pharmacy.
URL Obscuring	17%	Encoding a URL in hexadecimal, hiding the true URL with an @ sign, etc.
Domain Spoofing	50%	Using an invalid or fake domain in the from line.

The techniques used by spammers to overcome text classifiers are listed in Table 1. It shows that the most popular technique to confuse text classifier is Text Chaffing. It makes random strings of characters, series of words or unrelated sentences.

On that ground, Spam e-mail dataset was created using variety of text attributes consisting most common words and characters for spam emails.

The word appearance frequency in the email body text is expressed as percentage of words in the e-mail that match the WORD, from the total number of words in e-mail Equation. 1.

$$WORD_{freq} = 100 \cdot \frac{n_{WORD}}{N_{words}}, [\%] \qquad (1)$$

Where:

n_{WORD} – is number of times the WORD appears in the email.

N_{words} – is total number of any words in email.

The appearances of most common ASCII characters in the text is expressed as percentage of characters in the e-mail that match CHAR, from the total number of characters in e-mail Equation 2.

$$CHAR_{freq} = 100 \cdot \frac{n_{CHAR}}{N_{char}}, [\%] \qquad (2)$$

where

n_{CHAR} – is number of CHAR occurrences in the email.

N_{char} – is total number of characters in email.

In that manner collected 4602 exams, consisting of 57 attributes as input parameters and one attribute as output parameter, acquiring values (1,0), i.e. spam or legitimate email.

Artificial neural net for spam message classification

We applied a neural network (NN) approach to the classification of spam in this paper. The method employs attributes comprised from descriptive characteristics of the evasive patterns that spammers employ rather than the context or frequency of keywords in the messages. We will find out which ANN configuration will have the best performance and least error to desired output.

As the mathematical modeling platform we were using NeuroDimensios graphical neural network development tool NeuroSolutions. According to Neural Network theory, for static pattern classification the best performance shows the layered feedforward networks, called Multilayer Perceptrons (MLPs), typically trained with static backpropagation. Their main advantage is that they are easy to use, and that they can approximate any input/output map. The key disadvantages are that they train slowly, and require lots of training data.

ANN configuration and training

We had built different complexity MLP nets, to find out the most efficient structure, starting from simplest one – with 27 input nodes (giving 27 *word* and *character* attributes) , ending – with structure having 57 input nodes, and with different complexity of hidden layer for each case as also [7].

The generated data set was randomly split into three mixed groups: a training set (n=3220), a cross validation set (n=690), and a test set (n=690). For MLP training used training exams and for cross validation used cross validation exams.

For current modeling is used a log-sigmoid learning function keeping the ANN output to a continuous range between zero and one. It is more convenient as the dataset flags a spam email as being one and a non spam email as zero (values outside this range are unwanted).

The actual values of the mean squared error (MSE) using training and cross validation data set of the different structures MLP nets are showed in the following figures, and best training results of actual net structure are given in following tables.

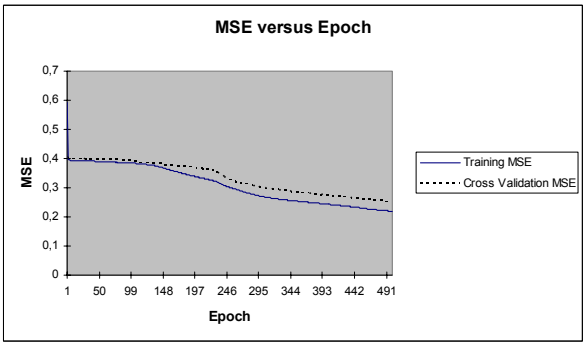

Fig. 1. MSE of training data set relation to epoch number of ANN having structure 27-20-1 (27 inputs, 20 nodes in hidden layer and 1 output node).

Table 2. Best networks 27-20-1 training results.

Best Networks	*Training*	*Cross Validation*
Epoch #	498	495
Minimum MSE	0,170491369	0,202818488
Final MSE	0,170892393	0,203013452

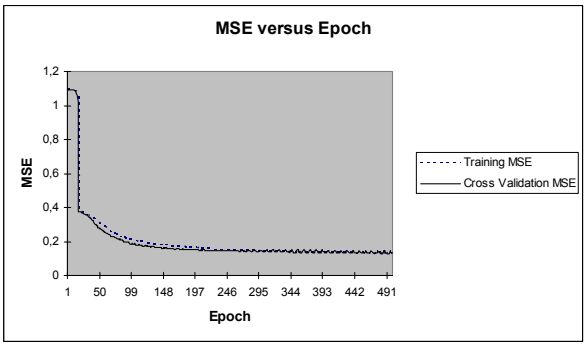

Fig. 2. MSE of training data set relation to epoch number of ANN having structure 41-10-1.

Table 3. Best networks 41-10-1 training results.

Best Networks	*Training*	*Cross Validation*
Epoch #	484	491
Minimum MSE	0,128083285	0,12947799
Final MSE	0,130594041	0,130826076

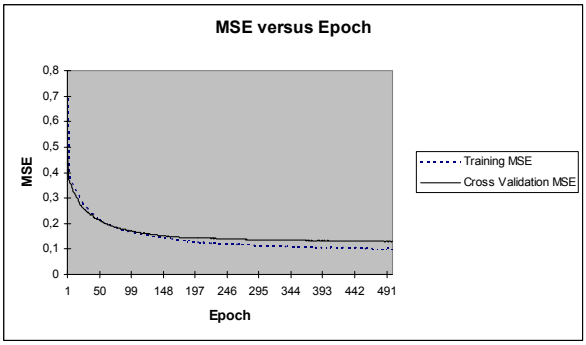

Fig. 3. MSE of training data set relation to epoch number of ANN having structure 57-20-1.

Table 4. Best networks 57-20-1 training results

Best Networks	*Training*	*Cross Validation*
Epoch #	496	498
Minimum MSE	0,097554186	0,128668806
Final MSE	0,098899114	0,129372457

From given tables it is obvious that after 500 training epoch the best result gives net, having structure 57-20-1 responding minimum MSE=0,098 and best cross validation data MSE value too.

Testing ANN

The trained ANNs during the testing stage is presented with the testing set (n=690), the dataset, which the ANN has not yet "seen" before. So the testing stage will show how many of the emails are correctly identified. This is done by comparing the actual result of the ANN to the target result contained in the testing set. This will be the basis for most of the analysis, optimizing the use of ANN.

The relative success of spam filtering techniques is determined by measuring of precision on the testing data sets. Spam precision is defined as the percentage of messages classified as spam that actually are spam. Likewise, legitimate precision is the percentage of messages classified as legitimate that are indeed legitimate and false parameters as also.

The recognition capability of the ANN as precision parameter is described by confusion matrix, giving the most actual parameters: *False positive* – Legitimate message classified as Spam and *False negative* – Spam classified as Legitimate e-mail.

In following tables presented testing results of each ANN structure.

Table 5. 27-20-1 ANN's structure Confusion Matrix and performance

Output / Desired	*Legitimate*	*Spam*	*Legitimate*	*Spam*
As legitimate	387	46	89,17 %	17,97 %
As Spam	47	210	10,83 %	82,03 %
Performance				
MSE	0,1046			

Table 6. 41-10-1 ANN's structure Confusion Matrix and performance

Output / Desired	*Legitimate*	*Spam*	*Legitimate*	*Spam*
As legitimate	399	36	94,55 %	13,43 %
As Spam	23	232	5,45 %	86,57 %
Performance				
MSE	0,0694			

Table 7. 57-20-1 ANN's structure Confusion Matrix and performance

Output / Desired	*Legitimate*	*Spam*	*Legitimate*	*Spam*
As legitimate	403	25	94,82 %	9,43 %
As Spam	22	240	5,18 %	90,57 %
Performance				
MSE	0,0597			

Comparing the results of confusion matrixes, we can consider that ANN which was trained using 57 email parameters, produced the lowest number of misclassifications, giving the lowest *False positive* = 5,18 % and *False negative* = 9,43 % values, for a total of 22+25 misclassified emails.

Conclusions

The recognition capability of the ANN is found to be good, but because of a low but nonzero false spam recognition rate (real messages erroneously classified as spam) the ANN is not suitable for use alone as a spam rejection tool. In fact any nonzero false positive spam detection rate is unacceptable, since the rejected email could be important message for recipient.

False identification of legitimate email is worse than receiving spam message, so the filter that yields false positive is not suitable. However there is no sense to increase the text parameters number, because the difference of classification precision considering false

positive value between 41 and 57 input nets is minimal.

The ANN, unlike the statistical classifier, is possible to train with additional input parameters related not only to text, but implementing pattern recognition also, especially when most today's spam messages are not even included with unsolicited text which allows quickly recognize spam using ANN identifying keywords, but uses graphical media as attachment to normal text e-mail.

Strategies that apply a combination of techniques, such as a NN with a whitelist, pattern recognition would yield better results.

References

1. **MessageLabs Ltd**, Spam review // Research report. – MessageLabs Ltd.; http://www.messagelabs.com
2. **Communications Regulatory Authority,** Lithuanian Internet users Inquest report, http://www.rrt.lt/index.php?-1158746316**;** last accessed February, 2006.
3. **Hauser. S**. 2003. Statistical Spam Filter Review. http://www.sofbot.com/article/Spam_review.html; last accessed January, 2006.
4. **Weiss, A.** Ending Spam's Free Ride // netWorker. – 2003. – No. 7(2). – P. 18–24.
5. **Graham, P.** 2002. A Plan for Spam. http://www.paulgraham.com/spam.html; last accessed January, 2006.
6. **Stolfo S., Li W., Hershkop S., Wang K.**, Detecting Viral Propagations Using Email Behavior Profiles // Research report. – Columbia University, 2003 .– 60 p.
7. **Principe C. J., Euliano R. N.**, Neural and Adaptive Systems: Fundamentals Through Simulations // John Wiley and Sons, 2000. – USA. – P. 350-400.
8. **Laurutis R., Puniškis D.**, Neural networks for computer virus epidemics recognition. // Electronics and Electrical Engineering. – Kaunas: Technologija, 2005. – No. 4(60). – P. 28– 32.
9. **Laurutis R.** Neural networks for data security // Electronics and Electrical Engineering – Kaunas: Technologija, 2003. – Nr. 4(46). – P. 61– 64.

Submitted for publication 2006 02 23

D. Puniškis, R. Laurutis, R. Dirmeikis. An Artificial Neural Nets for Spam e-mail Recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 5(69). –P. 73–76.

The volume of unsolicited commercial e-mail messages transmitted by the Internet has reached epidemic proportions nowadays. As a computer viruses spam is changing all time. Spam gets not only new forms every day but new exploit techniques also. The recognition capability of the ANN is found to be good, but because of low but nonzero false spam recognition the ANN is not suitable for use alone as a spam rejection tool. In fact, false identification of legitimate email is worse than receiving spam message, so the filter that yields false positive is not suitable. From other hand there is no sense to increase the text parameters number, because the difference of classification precision considering false positive value between 41 and 57 input nets is minimal. The ANN, unlike the statistical classifier, is possible to train with additional input parameters related not only to text, but implementing pattern recognition also, especially when most today's spam messages are not even included with unsolicited text which allows quickly recognize spam using ANN identifying keywords, but uses graphical media as attachment to normal text e-mail. Ill. 3, bibl. 9 (in English; summaries in English, Russian and Lithuanian).

Д. Пунишкис, Р. Лаурутис, Р. Дирмейкис. Нейронные сети для фильтрации спама в электронной почте // Электроника и электротехника. – Каунас: Технология, 2006. – № 5(69). – С. 73–76.

Доля спама в электронной почте неуклонно растет и, по оценкам экспертов, уже достигла эпидемический уровень. Как и вирусы, спам всё время меняется. Не только появляются всё новые и новые письма, но и постоянно расширяется арсенал приемов, используемых спамерами для обхода фильтров. Во время моделирования выяснено, что нейронная сеть способна с определенной четкостью идентифицировать спам письма, поэтому нейронная сеть должна использоваться в сочетании с другими технологиями, разрешающими повысить точность фильтра. Неверное распознавание нормального письма причисляя его к классе спам более вредно чем пропускание спама, и такой классификатор выдающий положительную ошибку является непригодным. С другой стороны, увеличивание параметров текста с 41 до 57 в отношении положительной ошибки дает незначительное улучшение и дальнейшее увеличение неизменит эффективность классификатора. На сегодняшний день спамерами используются приемы для обмана почтовых фильтров: случайные последовательности, замена и удвоение букв, текст "белым по белому" и самый сложный к правильном определении – в виде рекламных картинок, часто ссылкой на рекламный вебсайт. В таком случае необходимо дополнить пространство параметров обучения, способным обучить нейронную сеть анализировать текст предъявляемый в виде графики. Ил. 3, библ. 9 (на английском языке; рефераты на английском, русском и литовском яз.).

D. Puniškis, R. Laurutis, R. Dirmeikis. Neuroninių tinklų naudojimas nepageidaujamo elektroninio pašto žinutėms filtruoti // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2006.– Nr. 5(69). –P. 73–76.

Nepageidaujamo pašto telekomunikacijų tinkluose kiekvienais metais daugėja. Kaip ir kompiuteriniai virusai, komercinio turinio žinutės, siekdamos apeiti filtrus, nuolat kinta. Kiekvieną dieną atsiranda ne tik naujų turinio formų, bet ir tobulėja jų kūrimo ir platinimo strategija. Modeliuojant nustatyta, kad neuroniniai tinklai sugeba atskirti nepageidaujamą paštą nuo normalaus el. laiškų srauto, tačiau galimas tam tikras klaidingų sprendimų skaičius, todėl turi būti taikomas kartu su kitomis teksto ar vaizdo atpažinimo technologijomis, leidžiančiomis pagerinti filtro tikslumą. Normalų laišką klaidingai priskirti prie reklaminių yra blogiau, nei praleisti reklaminį laišką kaip normalų, todėl klasifikatorius, apskritai duodantis teigiamą klaidą, yra netinkamas. Kitą vertus, didinant mokymo parametrų skaičių nuo 41 iki 57, teigiamų klaidų sumažėja nedaug, todėl nauji teksto parametrai šiuo atžvilgiu nėra reikšmingi. Nepageidaujamo pašto siuntėjai stengiasi apeiti automatinius teksto atpažinimo filtrus naudodami įvairias žodžių ir simbolių kombinavimo metodikas. Viena iš naujausių – normalaus teksto pašto žinutėje pateikimas šalia prikabinant reklaminį paveikslėlį, paprastai susieta su nuoroda į reklaminį tinklalapį. Tokiu atveju būtina papildyti neuroninio tinklo mokymo parametrų aibę parametrais, įgalinančiais tinklą analizuoti ir atpažinti tekstą, pateiktą grafine forma. Ill. 3, bibl. 9 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).