Lecture Notes in Management Science (2014) Vol. **6**: 58–64
6th International Conference on Applied Operational Research, Proceedings
© Tadbir Operational Research Group Ltd. All rights reserved. www.tadbir.ca

ISSN 2008-0050 (Print), ISSN 1927-0097 (Online)

A generalized additive neural network application in information security

Tiny du Toit and Hennie Kruger

North-West University, Potchefstroom Campus, Potchefstroom, South Africa
{tiny.dutoit, hennie.kruger}@nwu.ac.za

Abstract. Traditionally spam has been considered as an inconvenience requiring workers to sift through and delete large numbers of e-mail messages per day. However, new developments and the Internet have dramatically transformed the world and over the last number of years a situation has been reached where inboxes have been flooded with unsolicited messages. This has caused spam to evolve into a serious security risk with prominent threats such as spreading of viruses, server problems, productivity threats, hacking and phishing etc. To combat these and other related threats, efficient security controls such as spam filters, should be implemented. In this paper the use of a Generalized Additive Neural Network (GANN), as a spam filter, is investigated. A GANN is a novel neural network implementation of a Generalized Additive Model and offers a number of advantages compared to neural networks in general. The performance of the GANN is assessed on three publicly available spam corpora and results, based on a specific classification performance measure, are presented. The results showed that the GANN classifier produces very accurate results and may outperform other techniques in the literature by a large margin.

Keywords: generalized additive neural network; information security risk; neural network; spam

Introduction

Undesired electronic messages have become a serious concern with spam comprising up to 88-90% of the total amount of e-mail messages (MAAWG, 2011). Several problems are caused by spam with some producing direct financial losses (Blanzieri and Bryl, 2008). More precisely, computational power, storage space and traffic are misused (Siponen and Stucke, 2006); additional mail must be sorted and looked through which irritates users and results in a loss of work productivity and time; many users claim spam violates their privacy rights (Siponen and Stucke, 2006); finally, legal problems are caused by spam which advertises pyramid schemes, pornography etc. (Moustakas *et al*., 2005). Various definitions of what spam (junk mail) is and how it differs from legitimate mail (genuine mail, ham or non-spam) can be found (Blanzieri & Bryl, 2008). Androutsopoulos *et al*. (2000a) characterizes spam as "unsolicited bulk e-mail".

Spam is no longer just considered as an invasive annoyance or a problem of convenience but it is regarded and accepted as an issue which poses a considerable security risk to enterprises. This view is due to the fact that spam is used, amongst other things, for spreading computer viruses and as a deceptive method of obtaining sensitive information. It has already been reported that about ninety percent of companies agreed that spam makes their companies more vulnerable to security threats (CNET News, 2004). More recent surveys and reports support this point of view (Jansson, 2011); (Deloitte Touche, 2011); (Ernst & Young, 2011). It has therefore become imperative to ensure that proper policies and controls are in place to mitigate the security risks associated with spam. One important control is the detection and management of spam messages. In this paper a Generalized Additive Neural Network (GANN) model is applied to three publicly available spam corpora to provide insight into the feasibility of using a GANN to filter spam messages.

The rest of the paper is organized as follows. First, the GANN architecture which is the neural network implementation of a Generalized Additive Model is discussed. Next, the three publicly available corpora are introduced. An experiment to classify incoming spam e-mail to determine the predictive accuracy of the GANN architecture is described. As preprocessing steps, vector representations of the messages are constructed and feature selection is performed. In addition, the GANN is compared to a Naïve Bayesian classifier as well as a Memory-based technique. Results obtained are analysed and some conclusions are presented in the last section.

The Generalized Additive Neural Network Architecture

Filtering is often used as a solution to the spam problem. To arrive at a spam filter, a decision function f must be obtained that automatically classifies a given e-mail message m as spam (S) or legitimate mail (L) (Blanzieri and Bryl, 2008):

$$f(m,\theta) = \left\{ \begin{array}{l} L \text{ if the message is regarded as spam} \\ S \text{ if the message is regarded as legitimate email} \end{array} \right\}$$

where m is the message to be classified, θ is a vector of parameters and S and L are labels assigned to the messages. Spam filters are usually based on machine learning classification techniques. The vector of parameters θ is the result of training the classifier on a pre-collected dataset:

$$\theta = \Theta(M),\ M = \{(m_1, y_1), (m_2, y_2), \cdots, (m_n, y_n)\}, y_i \in \{S, L\},$$

where $m_1, m_2, \ldots, m_n$ are messages collected previously, $y_1, y_2, \ldots, y_n$ are the corresponding labels, and Θ is the training function.

A Generalized Additive Model (GAM) is defined as

$$g_0^{-1}(E(y)) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_k(x_k),$$

where the expected target on the link scale is expressed as the sum of unspecified univariate functions (Wood, 2006); (Hastie and Tibshirani, 1990); (Hastie and Tibshirani, 1986). Each univariate function can be regarded as the effect of the corresponding input while holding the other inputs constant. When a GAM is implemented as a neural network it is called a Generalized Additive Neural Network (GANN) (Potts, 1999).

The main architecture of a GANN is comprised of a separate Multilayer Perceptron Neural Network (MLP) with a single hidden layer of h units for each input variable:

$$f_j(x_j) = w_{1j} \tanh(w_{01j} + w_{11j} x_j) + \cdots + w_{hj} \tanh(w_{ohj} + w_{1hj} x_j).$$

The individual bias terms of the outputs are incorporated into the overall bias β_0. Each individual univariate function contains $3h$ parameters, where h, the number of hidden neurons, could be different across inputs. This architecture can be extended to include an additional parameter for a direct connection (skip layer):

$$f_j(x_j) = w_{0j} x_j + w_{1j} \tanh(w_{01j} + w_{11j} x_j) + \cdots + w_{hj} \tanh(w_{ohj} + w_{1hj} x_j).$$

A backfitting algorithm is used by Hastie and Tibshirani (1986, 1990) to estimate the individual univariate functions f_j. Backfitting is not required for GANNs. Any method that is suitable for fitting more general MLPs can be utilized to simultaneously estimate the parameters of GANN models. The usual optimization and model complexity issues also apply to GANN models.

Currently, two algorithms exist to estimate GANN models. An interactive construction algorithm that makes use of visual diagnostics to determine the complexity of each univariate function was suggested by Potts (1999). When GANNs are constructed interactively, human judgment is required to interpret partial residual plots. For a large number of inputs this can become a daunting and time consuming task. Also, human judgment is subjective which may result in the creation of models that are suboptimal. Consequently, Du Toit (2006) developed an automated method based on the search for models using cross-validation or objective model selection criteria. With this approach, partial residual plots are used as a tool to provide insight into the models constructed and not primarily for model building. When given adequate time to evaluate candidate models, this best-first search technique is complete and optimal. Du Toit showed the algorithm is powerful, effective and produces results comparable to other non-linear model selection techniques found in the literature.

Prior to discussing the Naïve Bayesian classifier and the Memory-based classifier to which the results of the GANN will be compared, specific vector notation to represent e-mail messages will be given. In addition, the Ecue, Ling-Spam and PU1 spam collections and the preprocessing steps performed on the corpora are considered next.

Corpora Collection and Preprocessing

The Ecue collection (Delany *et al.*, 2006), Ling-Spam collection (Androutsopoulos *et al.*, 2000b) and PU1 corpus (Androutsopoulos *et al.*, 2000a) were used in the experiments.

As in Sahami *et al*. (1998) a vector representation $X = (x_1,x_2,...,x_n)$ is constructed for every message in the three corpora where $x_1,x_2,...,x_n$ denote the values of attributes $X_1,X_2,...,X_n$. These values are binary with $X_i = 1$ if some characteristic corresponding to X_i is present in the message and $X_i = 0$ otherwise. For the experiments, each attribute indicates whether a particular word (e.g. money) can be found in the message.

Feature selection is performed by ranking the candidate attributes by their Information Gain (IG) (Blanzieri and Bryl, 2008) values and choosing those attributes with the highest IG scores.

Naïve Bayesian and Memory-Based Classification

For the experiments, the GANN is compared to a standard Naïve Bayesian classifier (Han and Kamber, 2012) utilized by Androutsopoulos *et al*. (2000b) as well as a Memory-based technique implemented by the TiMBL system which was applied by Androutsopoulos *et al*. (2000b) to the Ling-Spam corpus. Different nearest neighbourhoods (1, 2, and 10) were chosen for the TiMBL classifier.

Classiffication Performance Measure

The total cost ratio (TCR) enables the performance of a filter to be easily compared to that of the baseline where no filter is present (Androutsopoulos *et al*., 2000b). Higher TCR values suggest better performance. When TCR < 1 it is better not to utilize the filter (baseline approach). An intuitive meaning of TCR can be obtained by assuming cost is proportional to wasted time. Therefore, TCR measures how much time is spend manually deleting spam messages when no filter is used compared to the time spend manually deleting spam messages that passed the filter plus time needed to recover from legitimate messages mistakenly blocked.

Experimental Results

In the experiments a parameter was used to denote the cost of misclassification. For this paper the parameter (λ) was set to 1 and the number of selected attributes by Information Gain ranged from 25 to 100 in steps of 25. With $\lambda = 1$ it is assumed the cost of misclassification is equal for the two types of errors (L classified as S or S classified as L). In all the experiments, 10-fold cross-validation were performed.

Figure 1 and Table 1 show the results obtained by the GANN model on the three corpora respectively. The GANN improves significantly on the baseline (TCR = 1.0) and produces very accurate results (high TCR values) on all three corpora. Moreover, the GANN outperforms the other two techniques by a large margin on the Ling-Spam corpus (Table 1). From Figure 1 it can be perceived that the GANN demonstrates a trend of highly accurate performance on spam corpora. Table 1 also shows the best results obtained by the GANN model on the Ecue and PU1 corpora respectively.

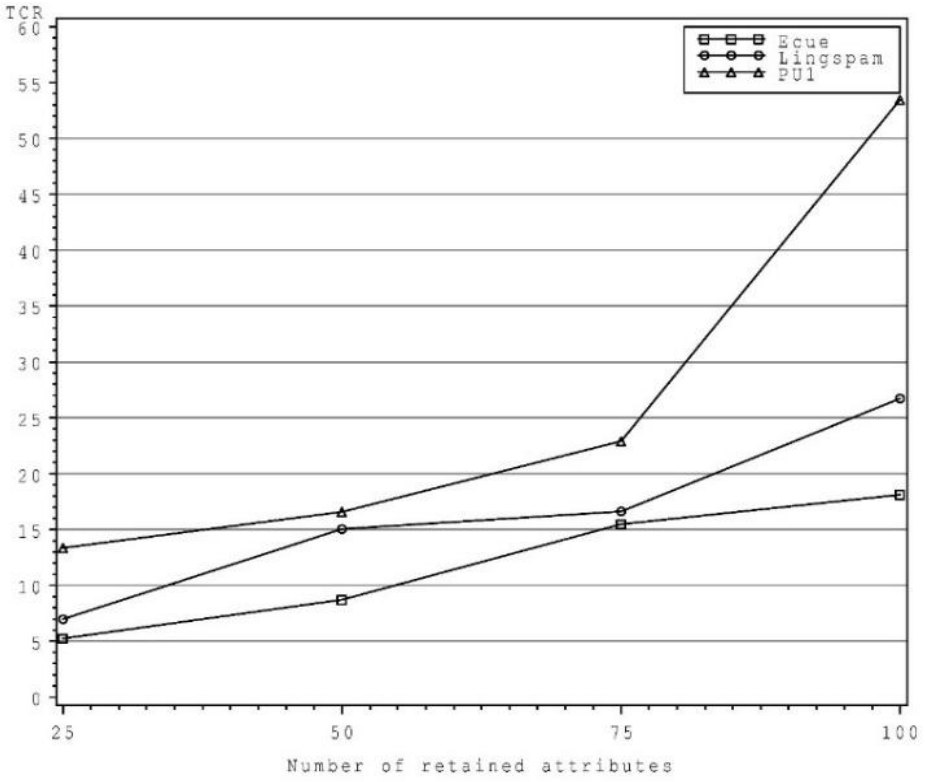

Figure 1. TCR scores obtained by the GANN on all three corpora with λ = 1.

Table 1. TCR scores obtained by the GANN compared to other filters.

Filter used	*Corpus*	λ	*# Attributes*	*TCR*
Naïve Bayesian	Ling-Spam	1	100	5.41
TiMBL(1)	Ling-Spam	1	50	5.35
TiMBL(2)	Ling-Spam	1	50	5.12
TiMBL(10)	Ling-Spam	1	100	1.52
GANN	Ling-Spam	1	100	26.71
GANN	Ecue	1	100	18.106
GANN	PU1	1	100	53.438
Baseline (no filter)	-	-	-	1

Conclusions

Modern technology and the Internet have transformed the world into something where constant communication is possible. All these new communication channels allow e-mail messages to be sent to anyone thousands of kilometres away. Unfortunately this freedom of communication can be exploited and unsolicited messages, or spam, are now accepted as an issue which poses a considerable risk to enterprises. Spam and its associated risks are considered as part of the dark side of the Internet (Kim *et al*., 2011) due to its illegal, unethical, or at least reprehensible elements.

In this paper the Generalized Additive Neural Network (GANN) was evaluated as a spam filter. The GANN is relatively unknown and has a number of favourable properties which makes it a suitable candidate for spam detection (Potts, 1999). These promising features motivated the application of GANNs to the domain of spam filtering. Periodically updating the GANN model automatically ensures the filter stays accurate. The GANN model was applied to three publicly available corpora and, based on a cost-sensitive measure, it proves to be very accurate. The findings in the paper suggest that the GANN model can be used successfully as an anti-spam filter and that significant results can be obtained.

References

Androutsopoulos, I, Koutsias, J, Chandrinos, KV, Spyropoulos, CD (2000a). An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval. ACM Press, pp 160-167

Androutsopoulos, I, Paliouras, G, Karkaletsis, V, Sakkis, G, Spyropoulos, CD, Stamatopoulos, P (2000b). Learning to filter spam e-mail: A comparison of a naive bayesian and a memory-based approach. In: Zaragoza, H, Gallinari, P, Rajman, M. (Eds.), Proceedings of the workshop `Machine Learning and Textual Information Access'. 4th European Conference on Principles and Practice of Knowledge Discovery in Databases (KDD-2000), Lyon, France, pp 1-13

Blanzieri, E, Bryl, A (2008). A survey of learning-based techniques of email spam filtering. Artificial Intelligence Review 29 (1), pp 63-92

CNET News (2004). Spam seen as security risk, (http://news.cnet.com/spam-seen-as-security-risk/2100-7355_3-5157275.html, accessed 14 March 2012)

Collins (2012). Collins English Dictionary, Complete and Unabridged 10th Edition, HarperCollins Publishers, (http://dictionary.reference.com/browse/machine learning, accessed 7 June 2013)

Delany, SJ, Cunningham, P, Smyth, B (2006). Ecue: A spam filter that uses machine leaming to track concept drift. In: ECAI. pp 627-

Deloitte Touche (2011). Raising the bar, TMT Global Security Study - Key Findings. Report published by Deloitte, 24p

Du Toit, JV (2006), Automated Construction of Generalized Additive Neural Networks for Predictive Data Mining, PhD thesis, North-West University, Potchefstroom Campus, South Africa

Ernst & Young (2011). Global information security survey, (http://www.ey.com/GL/en/Services/Advisory/2011-Global-Information-Security-Survey|Into-the-cloud-out-of-the-fog, accessed 16 March 2012)

Han, J, Kamber, M (2012). Data Mining: Concepts and Techniques, 3rd Edition. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann Publishers, San Francisco

Hastie TJ and Tibshirani RJ (1986). Generalized additive models. Statistical Science 1, 297-318

Hastie TJ and Tibshirani RJ (1990). Generalized Additive Models. Monographs on Statistics and Applied Probability, Vol. 43, Chapman and Hall, London

Jansson, K. (2011). A model for cultivating resistance to social engineering attacks. Master's thesis, Nelson Mandela Metropolitan University, Port Elizabeth, South Africa

Kim, W, Jeong, O-R, Kim, C, So, J (2011). The dark side of the Internet: Attacks, costs and responses. Information Systems 36, pp 675-705

MAAWG (2011). Messaging Anti-Abuse Working Group - Email metrics report, (http://www.maawg.org/sites/maawg/ les/news/MAAWG_2011_Q1Q2Q3_Metrics_Report_15.pdf, accessed 1 March 2013)

Moustakas, E, Ranganathan, C, Duquenoy, P (2005). Combating spam through legislation: a comparative analysis of US and European approaches. In: Proceedings of the Second Conference on Email and Anti-Spam

Potts, WJE (1999). Generalized additive neural networks. In: KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA, pp 194-200

Sahami, M, Dumais, S, Heckerman, D, Horvitz, E (1998). A bayesian approach to filtering junk e-mail. Tech. Rep. WS-98-05, Learning for Text Categorization Papers from the AAAI Workshop, Madison Wisconsin

Siponen, M, Stucke, C (2006). Effective anti-spam strategies in companies: An international study. In: Proceedings of the 39th Annual Hawaii International Conference on System Sciences - Volume 06. HICSS '06. IEEE Computer Society, Washington, DC, USA, pp 127.3-

Wood, SN (2006). Generalized Additive Models: An introduction with R. Texts in Statistical Science. Chapman & Hall/CRC, London