Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Engineering

Elixir Comp. Engg. 77 (2014) 28908-28911

ISSN: 2229-712X

Spam detection by ANFIS with feature selection by GA

Masoumeh Esmaeili[1,*] and Morteza Zahedi[2]

[1]Department of Computer Engineering & IT, Shahrood University of Technology, Shahrood, Iran.

[2]Faculty of Computer Engineering & IT, Shahrood University of Technology, Shahrood, Iran.

ARTICLE INFO

Article history:

Received: 26 June 2014;

Received in revised form:

19 November 2014;

Accepted: 29 November 2014;

Keywords

Spam email, ANFIS, Fuzzy logic, Text classification, Spamicity.

ABSTRACT

Spam is the sending unwanted e-mail messages frequently with different contents, in large quantities to an indiscriminate set of recipients, and often proselytes a service or a website. Many intelligent systems have been developed for detecting spam emails, but many of them don't have enough speed. In this paper, a fuzzy spam detection system in text classification mode is described that has been implemented in MATLAB. Because of the ANFIS uses the approximation capability of FIS and ANN as adaptive, it acts simple and powerful. In the proposed method, first extractor starts to extract all the tokens in the body of all emails. Genetic algorithm (GA) is then applied, to select the appropriate features of the tokens. These features are saved in a dictionary. Then ANFIS uses this dictionary for classifying emails. In this project, ANFIS has three inputs and one output. For obtaining ANFIS' inputs, calculate a spamicity for each token. This criterion shows the rate of dangerous of each token. Then tokens of each email are classified into three categories, based on the amounts of their spamicity. Counts of tokens in each category, are three inputs to ANFIS system. ANFIS' output determines that each email is spam or not.

Introduction

Spam is an email with irrelevant contents that is sent automatically with different contents for many different users and often proselytes a service or website [1]. Spam decreases the reliability of e-mail [2]. Even aware users aren't comfortable from these unwanted emails. Spam is undesirable because it eats up resources like disk space and user time. Several methods for classifying emails exist [3]-[5], but each has certain weaknesses and some of them don't have enough speed. The main idea in spam detection is defining an appropriate threshold between spam and non-spam emails that minimize misclassification [6]. We have implemented a fuzzy spam detection system in text classification mode, which uses adaptive neuro-fuzzy inference system (ANFIS) for classifying emails. Feature selection methods can be classified essentially into wrappers, filters and embedded techniques [7],[8]. In this paper, GA is used as feature selection in order to increase speed of classification, and then selected features are used for classifying emails using ANFIS in the framework of embedded methods [9], which is named as GA-ANFIS.

ANFIS uses the approximation capability of FIS and ANN as adaptive, and also acts simple and powerful. In other words, not only does it have good learning capability, but it can be also universally approximate well. Moreover, the training of ANFIS is fast and it can generally converge from small datasets. These attractive properties are suitable for choosing ANFIS as classifier for classifying emails.

Adaptive-Network-Based Fuzzy Inference System

ANFIS is a class of adaptive networks that are functionally equivalent to fuzzy inference systems. ANFIS represents Sugeno fuzzy inference model (SFIM) that was proposed by Jang in 1993 [10], [11]. The fuzzy inference systems (FIS) and multi-layer perceptrons (MLP) [12], [13] are special examples in generic calculation studies of adaptive networks [14]. By having dataset, the ANFIS model employs the neural network training procedure to adjust the membership function parameters with using back propagation algorithm or other similar optimization methods[10], [15], [16], [17]. Block diagram of fuzzy inference system has been shown in Fig.1.

To explain the structure of ANFIS, assume that if our fuzzy Inference system had two inputs x and y and had one output f, and rule base has two following if-then rules [18]:

Rules 1: If (x is A_1) and (y is B_1) then ($f_1=\alpha_1 x+\beta_1 y+\gamma_1$)

Rules 2: If (x is A_2) and (y is B_2) then ($f_2=\alpha_2 x+\beta_2 y+\gamma_2$) (1)

In these formula the concise forms, x and y are the inputs, A1, A2, B1 and B2 are the fuzzy sets determined during the network training procedure, f1 and f2 are outputs, α1, α2, β1, β2, γ1 and γ2 are linear parameters, which are also determined during the network training procedure [19], [15].

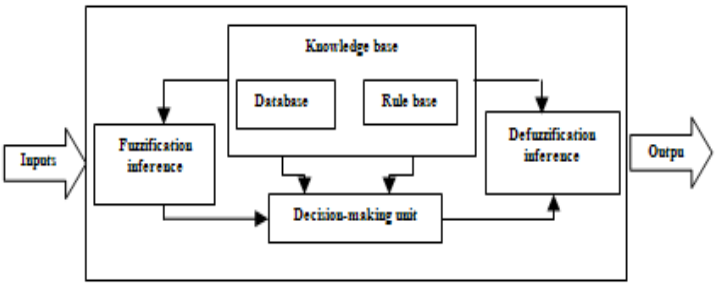

Figure 1 Block diagram of fuzzy inference system

The ANFIS structure consists of a combination of three fixed layers and two adaptive layers. The adaptive layers are the first and the fourth layer. In the first layer, there are three modifiable parameters (ai, bi and ci) related to the input membership function. In the fourth layer, there are also three modifiable parameters (pi, qi and ri), pertaining to the first-order polynomial of SFIM. These adaptive parameters can obtain a

Tele:

E-mail addresses: esmaeili.m2011@gmail.com

desirable value by learning algorithm. In this study, the learning algorithm of ANFIS is based on the hybrid learning algorithm. In the fifth layer, there is only one single fixed node, which is labeled as R. The node performs the summation of all the incoming nodes, which represents the defuzzification procedure [10], [15].

Feature Extraction

For implementing the system, we train system with 90 emails that contain 40 spam and 50 non spam emails. First extract all tokens from body of all emails and save them with their frequencies in spam and non spam emails, in three separate columns of a dictionary. Then, for each token, based on the frequencies, calculate a measure called spamicity, and save it in forth column of the dictionary. Finally use ANFIS to determine whether each email is spam or not. The block diagram of proposed method has been shown in Fig.2. The system structure of ANFIS is shown in Fig.3.

After making the original dictionary from training emails, additionally, make a local dictionary for each test email, that has three columns: first column contains all extracted tokens from that email, second column contains the frequency of each token in test emails and third column contains spamicity of each token that will be calculated based on the frequency of them in original dictionary. Some of tokens are ignored such as dot, comma, blank, etc. For applying ANFIS in this system, first should calculate spamicity of tokens that exist in original dictionary. This criterion will be saved in the fourth column of the dictionary.

The spamicity of token i from dictionary, is calculated like formula 2. This formula uses the second and third columns of the original dictionary, which show the frequency of tokens in spam and non spam emails, respectively.

$$Spamicity_i = \frac{frqcy\_token_i\_Spam}{frqcy\_token_i\_Spam + frqcy\_token_i\_NonSpam + 1} \quad (2)$$

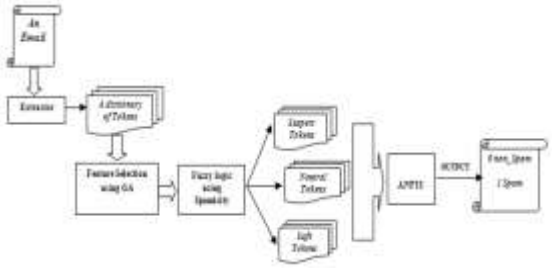

Fig 2. Block Diagram of Proposed Method

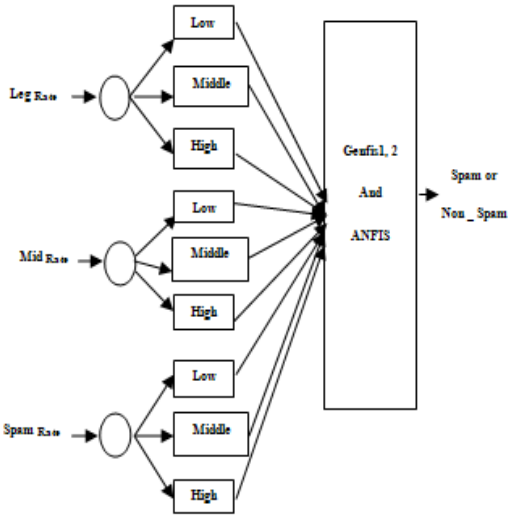

Fig 3. Block diagram of spam detection using ANFIS

Feature Selection

The keyword-based methods usually consider thousands of keywords as features for obtaining satisfactory results [20]-[22], [23]. But the original dictionary contains useful keywords and additionally irrelevant keywords which play no important role in classification. A proper selection of features can actually improves the classifying and generalizing ability of the classifier. For choosing such features, there are different feature selection methods. So, we use genetic algorithm for selecting best combination of features in order to increase the speed and performance of the system.

Genetic Algorithm

Genetic algorithm is an iterative method and consists of four main steps: producing initial population, evaluation, reproduction (consists of selection & crossover), and mutation. In the first step, a population of strings called chromosome, should be generated. Each chromosome, selects a singular combination of features from dictionary, randomly. In the second step, chromosomes must be evaluated. For this, classify test emails based on each chromosome, separately. The success rate of the classifier will be saved as fitness value of it. In selection step, half of the best chromosomes are selected based on their fitness values and they will be combined in crossover step. With combining the best of chromosomes, in several epoch s iteratively, a new population will be created in each epoch that is smarter than previous [24]. The new population is replaced with previous population and is used in the next iteration of the algorithm. Commonly, the algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been obtained for that population. Finally, select the best chromosome from the last population and create a new dictionary based on it. The features of the new dictionary are less than the previous, but the new dictionary is more efficient in time and performance [25]. Mutation operation is performed at the end of each epoch. In this operation, one of the chromosomes will be randomly selected and one or more of its genes are modified, randomly. So the modified value of that chromosome, may be does not exist in any of its parents [24].

Classification by ANFIS

Calculating ANFIS inputs

For training ANFIS system, inputs should be determined. ANFIS system that is used for spam detection has three inputs. First input is LegRate that shows the number of tokens with low spamicity, second input is MidRate that shows the number of tokens with middle spamicity and third input is SpamRate that shows the number of tokens with high spamicity. So, in spam emails spamRate should be greater than LegRate. The logic of this method is that the spam emails have many common tokens that rarely appear in non spam emails, for non spam emails it is same. Tokens with neutral spamicity, are those tokens that often appear routinely in any email, whether spam or non spam.

For calculating inputs, make three categories from tokens in body of each email. These categories include three types of tokens that are different in value of spamicity. First legitimate category that shows the tokens with low spamicity, second is neutral category that shows the tokens with middle spamicity, and third, spam category that shows the tokens with high spamicity. So, counts of tokens in each category, are three inputs of ANFIS system, namely LegRate, MidRate and SpamRate. To obtain more accurate inputs, they must be normalized. So divide them into total number of them (formula 3, 4 and 5). There are three levels of inputs: high, middle and low. Inputs and output are shown in table 1.

$$input1 = LegRate = \frac{leg\_rate}{leg\_rate + mid\_rate + spam\_rate} \quad (3)$$

$$input2 = MidRate = \frac{mid\_rate}{leg\_rate + mid\_rate + spam\_rate} \quad (4)$$

$$input3 = SpamRate = \frac{spam\_rate}{leg\_rate + mid\_rate + spam\_rate} \quad (5)$$

Table 1: Inputs and output of neuro-fuzzy spam detection system

NAME	OUTPUT	INPUT1	INPUT2	INPUT3
TYPE	Result	LegRate	MidRate	SpamRate
MAIN AMOUNTS	Spam Non_spam	High Mid Low	High Mid Low	High Mid Low
ALTERNATIVE AMOUNT	{0,1}	{1,2,3}	{1,2,3}	{1,2,3}

Creating initial fuzzy system

For creating initial fuzzy system, two different methods are implemented in this paper. First is Genfis1 method and second is Genfis2 method.

Genfis1

Genfis1 with using Grid Partition method on the training data, creates an initial fuzzy system Sugeno(FIS) with one output, for learning ANFIS. The obtained FIS, is used for initial conditions of ANFIS. For initializing, Genfis1considers bell-shaped membership functions, and rules will be produced linearly. The membership functions for each input are drawn, that shown in Fig.4. part (a).

Genfis2

Genfis2 with using Fuzzy Subtractive Clustering method on the learning data, creates an initial fuzzy inference system (FIS). Another different is that we should give the input and output data distinguishably. This method, clusters data before creating FIS. Clustering data is done with extracting a set of rules that determine the behavior input data. In procedure extracting data, first use 'Subclust' function to determine the number of rules and membership functions. Then with using LLSE method, determine the last of the rules [15]. The number of rules that are created, are less than previous method. The membership functions of initial FIS that are created using Genfis2, are shown in Fig.4. part (b).

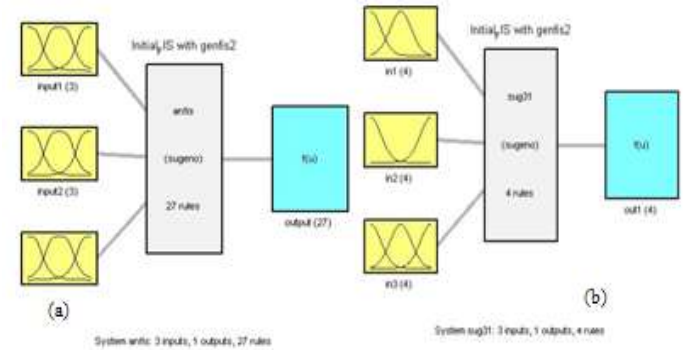

Fig.4. membership functions of initial FIS with number of rules, a) Genfis1 b) Genfis2

Evaluating to other parameters for learning

Important parameters such as epoch and tolerance should be determined before training. The value of epoch in Genfis1 and Genfis2 has been selected 15.

Starting learning step

In the next step, use ANFIS for learning system. So, for both of the obtained FIS s, use ANFIS command and for both states tolerance is given as output. ANFIS command uses a hybrid learning algorithm to determine FIS membership function parameters. A combination of LSE methods and BP gradient method, are used to train FIS membership function for modeling a set of input-output data. Membership functions and obtained FIS s, for both states shown in Fig. 4.

Learning error

We used 20 emails for testing system that contain 12 non-spam emails and 8 spam emails. Non_ spam emails are from email 1 to 12 that are shown with 0 and spam emails are from email 13 to 20 that are shown with 1. Fig.5 part (a) and (b) shows the classification results of Genfis1 method using GA and without GA respectively. And Fig.5 part (c) and (d) shows the classification results of Genfis2 method using GA and without GA respectively. As you can see, the blue diamond-shaped points and blue plus shaped points are results of ANFIS classification and red spots, show real classes of test emails. Non-compliance between these two types of points, are error in classification. The error classification rate without using GA is 35% in both Genfis1 and Genfis2, but with using GA, it is reduced to 28%.

Efficiency analyst

Comparing the two methods Genfis1 and Genfis2, Genfis2 is better and more optimum from Genfis1, because of having lower RMSE error. But Genfis2 usually is slower. As you can see in table2, with using GA, the error rate is 28%, which was declined 7%.

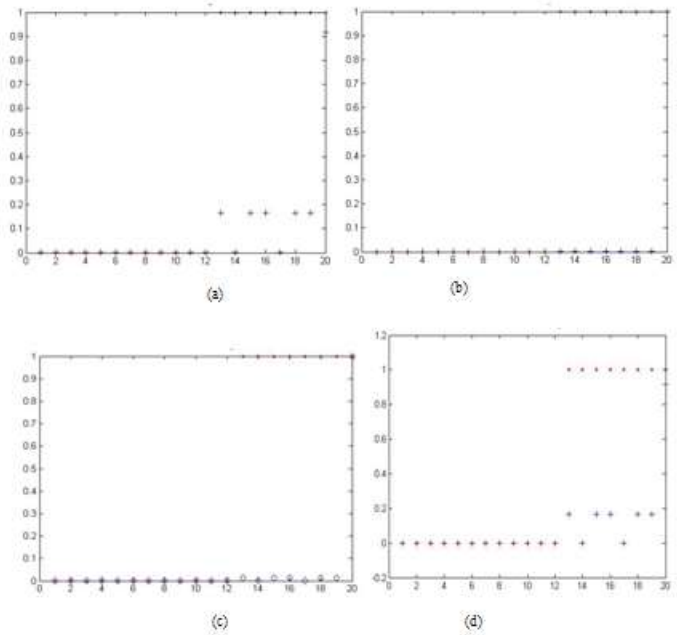

Fig 5. Classification results on test data, (a) Genfis1 using GA method, (b) Genfis1 without GA method, (c) Genfis2 using GA method, (d) Genfis2 without GA method

Conclusion

In this paper, an implementation for a neuro-fuzzy program for spam detection system was represented. This program first extracts all tokens from body of emails, and saves them with their frequencies in spam and non spam emails, in three columns of a dictionary. Then calculates a spamicity for all tokens and save them in forth column of the dictionary. The spamicity criterion shows dangerous rate of tokens. In order to more accurately categorize emails, genetic algorithm is used to select best combination of features. This method reduces useless features and increases the accuracy and speed of methods. For classifying emails we used ANFIS. It would be a good choice among other classifier methods, because combines the advantages of neural networks with that of FIS. For classifying emails, first use Genfis1 and Genfis2 to obtain membership functions of inputs, then use ANFIS on these initial FIS s to classify emails. The results show that Genfis1 is faster than Genfis2, but it is less accurate.

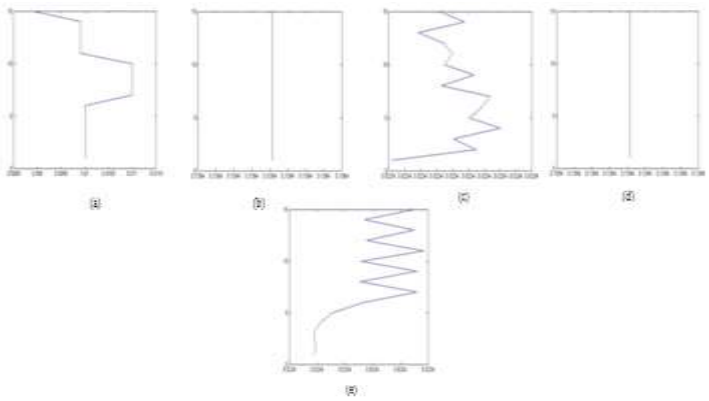

Fig. 6 step size and error rate of Genfis1 and Genfis2, (a) Step Size of Genfis1 and Genfis2, (b) Train Error of Genfis1, (c) Test Error of Genfis1, (d) Train Error of Genfis2, (e) Test Error of Genfis2

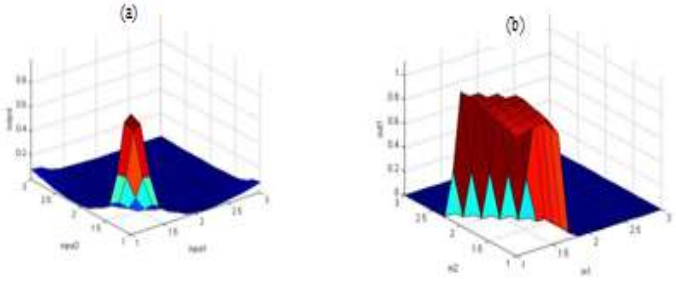

Fig.7 (a) The surface of final FIS, (a) Genfis1, (b) Genfis2

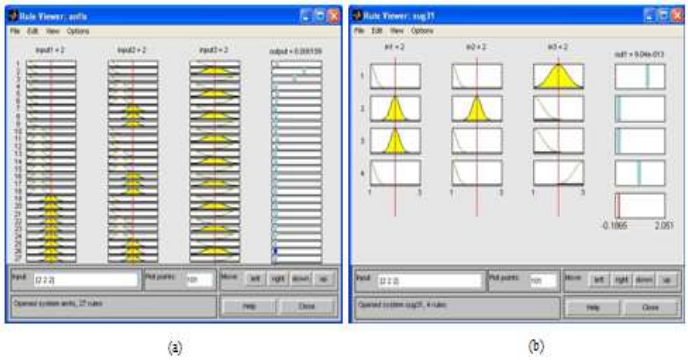

Fig 8 the rules and defuzzyfied of them in initial ANFIS, (a) Genfis1, (b) Genfis2

References

[1] Ch.H. Wu ,"Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks", ELSEVIER Expert Systems with Applications, Volume 36, Issue 3, Part 1, Pages 4321–4330, 2009.

[2] B. Hoanca, "How good are our weapons in the spam wars?", IEEE Technology and Society Magazine, 25(1), 22–30, 2006.

[3] W. Cohen, "Learning rules that classify e-mail", AAAI Spring Symposium on Machine Learning in Information Access, 1996.

[4] M. Sahami, S. Dumais, D. Heckerman, E. Horvitz. , "A Bayesian approach to filtering junk e-mail", In AAAI'98 Wkshp. Learning for Text Categorization, Madison, WI, July 27, 1998.

[5] H. Drucker et al ,"Support Vector Machines for Spam Categorization", IEEE TRANSACTIONS ON NEURAL NETWORKS, 10(5):1048-1054, 1999.

[6] A.H. Mohammada, R.A. Zitarb, " Application of genetic optimized artificial immune system and neural networks in spamdetection", Applied Soft Computing Volume 11, Issue 4, Pages 3827–3845 ,2011.

[7] R. Kohavi, G.H. John, "Wrappers for feature subset selection. Artificial Intelligence", 273-324, 1997.

[8] I. Guyon, A. Elissee, "An introduction to variable and feature selection", Journal of Machine Learning Research, 1157-1182, 2003.

[9] H. Wang, Y. Yu, Zh. Liu, "SVM Classifier Incorporating Feature Selection Using GA for Spam Detection", Dept. of computer science, Tianjin University of Technology, 300191, Tianjin, China.

[10] R. Jang, "Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence", Prentice Hall, 1996.

[11] R. Jang, "Adaptive Network- Based Fuzzy Inference Syste", IEEE Transactions On Systems, Man, and Cybernetics, Vol. 23, NO. 3, 1993.

[12] K. Tretyakov, "Machine Learning Techniques in Spam Filtering", Data Mining Problem-oriented Seminar, MTAT.03.177, pp. 60-79, 2004.

[13] T.S. Guzella , W.M. Caminhas, "A review of machine learning approaches to Spam filtering", IEEE Expert Systems with Applications 36 10206–10222, 2009.

[14] Ö. Deperlioğlua, U. Ergünb, G.E. Güraksınc, "Design of ANFIS Controller for DC-DC Step-Down Converter", Deperlioğlu, Ergün ve Güraksın/ AKÜ Fen Bilimleri Dergisi, 17-29, 2010.

[15] J.D. Wu, Ch. Hsu, H.CH Chen, "An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference", Expert Systems with Applications, Vol 36, 7809 – 7817, 2009.

[16] N. Nariman-Zadeh, "Design of ANFIS networks using hybrid genetic and SVD methods for the modeling of explosive cutting process", Journal of Materials Processing Technology, vol. 155, pp.1415- 1421, 2006.

[17] P. Sivarao, N.S.M. Brevern, El-Tayeb, V.C.V. engkatesh ,"ANFIS Modeling of Laser Machining Responses by Specially Developed Graphical User Interface", International Journal of Mechanical & Mechatronics Engineering IJMME Vol: 9 No: 9.

[18] K.C. Raveendranathan, M.Harisankar, "A New Class of ANFIS based Channel Equalizers for Mobile Communication Systems", IJSSST, Vol.11, No.1.

[19] L.A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes", IEEE Transactions on Systems, Man and Cybernetics, Vol.SMC-3, No.2, pp.28–44, 1973.

[20] E. Jiang, "Learning to semantically classify email messages", In *Proceedings of the international conference on intelligent computing*, pp. 700–711, *2006*.

[21] I. Kanaris, K. Kanaris,E. Stamatatos, "Spam detection using character n-grams", In G. Antoniou et al. (Eds.), Hellenic conference on artificial intelligence (pp. 95–104), 2006.

[22] R. Pampapathi, B. Mirkin, M. Levene, "A suffix tree approach to anti-spam email filtering Machine Learning, 65 , pp. 309–338, 2006.

[23] J. Androutsopoulos, K. Koutsias, C. Chandrinos, Spyropoulos, " An experimental comparison of naïve Bayesian and keyword-based anti-spam filtering with personal e-mail messages", Proceedings of SIGIR. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1-58113-226-3 (2000), pp. 160–167, 2000.

[24] M. Justin, "Filtering Spam With Spam Assassin", HEANet Annual Conference, 2002.

[25] G. Harik, F. Lobo, M. Kaufmann, "A Parameter-Less Genetic Algorithm. Proceedings of the Genetic and Evolutionary Computation", Conference, 258-265, 1999.