Neural Recognition and Genetic Features Selection for Robust Detection of E-Mail Spam

Dimitris Gavrilis[1], Ioannis G. Tsoulos[2], and Evangelos Dermatas[1]

[1] Electrical & Computer Engineering, University of Patras
gavrilis@upatras.gr, dermatas@george.wcl2.ee.upatras.gr
[2] Computer Science Department, University of Ioannina
sheridan@cs.uoi.gr

Abstract. In this paper a method for feature selection and classification of email spam messages is presented. The selection of features is performed in two steps: The selection is performed by measuring their entropy and a fine-tuning selection is implemented using a genetic algorithm. In the classification process, a Radial Basis Function Network is used to ensure robust classification rate even in case of complex cluster structure. The proposed method shows that, when using a two-level feature selection, a better accuracy is achieved than using one-stage selection. Also, the use of a lemmatizer or a stop-word list gives minimal classification improvement. The proposed method achieves 96-97% average accuracy when using only 20 features out of 15000.

1 Introduction and Related Work

The electronic mail is a crucial Internet service and millions of messages are sent every day. The flood of the user's mailboxes with unsolicited emails, a problem known as spam, consumes network bandwidth and can be seen as a Denial of Service attack. Many methods have been proposed to countermeasure spam but most of them do not employ machine learning but instead they use blacklists of known spammers, or dictionaries with phrase patterns usually found in spam messages. Other techniques require the users to manually identify and mark the spam messages in order to create personalized rules for each user.

Recent advances in text categorization (TC) have made possible the recognition of spam messages through machine learning techniques. The two key features that must be addressed in the TC problem are the feature selection and the classifier. The feature selection process is crucial and can improve performance because text contains a very large number of features (more than 20000 in an average in length corpus). In order to solve this problem, techniques like Singular Value Decomposition [7], term weighting based on text statistics [3] and latent semantic analysis (LSA) [8] have been used. The Vector Space Model (VSM) [3] and the nearest neighbor classifier [4] are primarily used as classifiers for document classification problems.

A simple k-NN classifier was used by Sakkis and Androutsopoulos [1], which is compared to a Naïve Bayes filter. The Information Gain (IG) is used to select the features. In the k-NN approach, the recall rate reaches 59.91% (using 600 features) while in the Bayesian approach is 63.67% (using 300 features).

G. Antoniou et al. (Eds.): SETN 2006, LNAI 3955, pp. 498 – 501, 2006.

The message preprocessing is another crucial factor in text categorization problems. Four different variations have been explored. A raw dataset (BARE), a dataset with its common words removed (STOP), a dataset with its words converted to their base form using a lemmatizer (LEMM) and a combination of the last two (LEMM-STOP).The authors in [2] explore the performance of a Naïve Bayes classifier on the four different datasets. Their results show that the LEMM-STOP gives better accuracy than the other approaches.

In this paper a genetic approach to the features selection problem for robust neural classification of e-mail spam is described and evaluated. In a two stages procedure, the initial set of a huge number of features are reduced using natural language tools, and selected the most stochastically important by measuring the features entropy. A fine-tuning selection is applied with the aid of a genetic algorithm.

In the following sections, the proposed method is described later and the corpus used for benchmark is presented. Finally, the experimental results are shown and the conclusions are given in the last section.

2 Feature Selection and Classification of E-Mail Spam

The method proposed in this paper consists of a feature selection process and a neural classifier. The feature selection process, in contrary with traditional methods, is performed in two levels. It has already been found that the proposed features selection outperforms other methods in text categorization, and especially in scientific abstract categorization [6].

The feature selection approach consists of two-level procedure. In the first level, the features are assigned a rank according to their entropy:

$$H(x) = -\sum_{y_i \in Y} p(y_i \mid x) \log p(y_i \mid x), \quad \text{i=1,2.} \tag{1}$$

where y is the class (legitimate or spam) and x is each feature. From the original set of *M* features, a subset is defined by selecting *N* features based on their entropy rank.

In the second level, the subset of *N* is scaled down to *K* features by genetic selection. A genetic algorithm using tournament selection, detects the best *K* features that are to be used for classification. The genetic optimization function maximizes the correct classification rate using a RBF network. Therefore, the proposed fine-selection of features is optimal.

For the classification task, a neural classifier is used, and more specifically an RBF network. Although neural networks are very good classifiers, are uncommon in text categorization problems since they are problematic with high dimensionality. In this paper it is shown that if a proper set of features is obtained ($K \approx 20$), excellent accuracy can be achieved. The training process of the RBF network is generally fast and that is the main reason that it is used, since the combination of genetic selection with a neural classifier is a very time consuming training process.

3 Benchmark Corpus

The corpus used for training and testing in this paper is the PU1 corpus [5] which is available at (http://www.aueb.gr/users/ion/data/pu1_encoded.tar.gz). The PU1 corpus has been created from a normal user for a period of 36 months. It contains 1099 english messages, 618 of them are legitimate and 481 are spam, and consists of 4 subsets: a raw set, a raw set with a stop-word list used, a set with the words lemmatized, a set with the words lemmatized and with a stop-word list used. The 4 sets are split into 10 groups and used to perform 10-fold cross validation in all experiments presented below. The corpus is also pre-processed and three other variations are created: the STOP, LEMM and LEMM-STOP corpuses. These variations are described in [5].

4 Experimental Results

The genetic algorithm in all experiments runs for 100 generations and has a 0.05% mutation rate and a 0.95% crossover rate. The RBF network has 6 hidden neurons and 1 neuron in the output layer. The k-means algorithm is used to derive the synaptic weights in the hidden layer.

The experimental results shown in Table 1, give the average error of the 10-fold validation for each dataset when all features were used for genetic selection classification (M=N=approximately 16000). The results show that, when the LEMM-STOP dataset is used, better accuracy is achieved.

Table 1. Mean classification error for each corpus when only the fine-selection process is used (genetic selection)

Corpus	Number of Features (M)	Average Error (%)
BARE	16000	9.81
STOP	16000	9.63
LEMM	16000	8.71
LEMM-STOP	16000	6.79

Table 2. Mean classification error for the LEMM-STOP dataset when both feature selection steps are used

Corpus	Number of Features (N)	Average Error (%)
LEMM-STOP	3000	5.44
LEMM-STOP	6000	3.27
LEMM-STOP	8000	4.71

When the the complete features selection method is used, the genetic selection uses a subset of the original features. In Table 2, the mean classification error for the LEMM-STOP dataset is shown using the complete version of the features selection method. The experiments are completed only on the LEMM-STOP dataset, since this

corpus gives lower classification error than the other three. It is obvious that the mean classification error drops significantly when only the genetic selection is used (Table 1). When a large number of features is selected, the classification error increases, because the search space of the genetic algorithm grows significantly. On the other hand, when keeping a small number of features, information helpful in the recognition process may be lost thus high classification error is measured. The use of a medium-size feature size (about 40-50% of the original) is found to give a low classification error. Although the results are promising, when the same method is applied to a classical text categorization problem, involving scientific paper abstracts ([6]), it performs significantly better. The smaller length in the email content can account for this divergence in performance.

5 Conclusions and Future Work

In this paper a two-step feature selection method is presented that uses term entropy to select a subset of the original features in the first step and genetic selection in the second step. An RBF network is used for the classification with 20 features as inputs producing a 3.27% classification error. Also, the use of a stop-word list and a lemmatizer is found to improve classification accuracy in the spam recognition problem. In future work, a larger corpus will be used and the performance of other classifiers will be explored.

References

1. G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C.D. Spyropoulos, and P. Stamatopoulos: A memory-based approach to anti-spam filtering for mailing lists. In Information Retrieval, Vol. 6. Kluwer (2003) 49-73.
2. Androutsopoulos, J. Koutsias, K. V. Chandrinos, G. Paliouras, and C. D. Spyropoulos: An Evaluation of Naive Bayesian Anti-Spam Filtering. In Proc. of the workshop on Machine Learning in the New Information Age (2000).
3. Dik L. Lee, Huei Chuang, Kent Seamons: Document Ranking and the Vector-Space Model. IEEE Software, Vol. 14 (1997) 67-75.
4. Michael Steinbach, George Karypis, Vipin Kumar: A Comparison of Document Clustering Techniques. KDD Workshop on Text Mining (2000).
5. E. Michelakis, I. Androutsopoulos, G. Paliouras, G. Sakkis and P. Stamatopoulos: Filtron: A Learning-Based Anti-Spam Filter. Proc. of the 1st Conference on Email and Anti-Spam (2004).
6. Dimitris Gavrilis, Ioannis Tsoulos and Evangelos Dermatas: Stochastic Classification of Scientific Abstracts. Proceedings of the 6th Speech and Computer Conference, Patra (2005).
7. John M. Pierre: On the Automated Classification of Web Sites. Linkoping Electronic Articles in Computer and Information Science, Vol. 6 (2001).
8. S. Deerwester, S. Dumais, G. Furnas, T. Landauer and R. Harshman: Indexing by latent semantic analysis. Journal of the American Society for Information Science, Vol. 46 (1990) 391-407.