ISSN:0975-9646

Finding Template Mails from Spam Corpus Using Genetic Algorithm and K-Means Algorithm

Liny Varghese, Supriya M.H, K. Poulose Jacob

Cochin University of Science and Technology,Cochin,
Kerala, India

Abstract: Spammers are using email templates for sending spam. To promote some service or product, they send mails by creating templates and merging the details of receivers with the template. Since they are using templates, similarities can be found among mails and spam detection softwares can easily ignore the forthcoming spam mails. Our objective is to identify the template mails from the whole corpora of training set emails to make the filtering process faster. In this paper we propose how supervised Genetic algorithm and K-Means algorithm can be used to generate the best population, which in turn used for spam classification.

Keywords: Spam, Genetic algorithm, K-Means algorithm, Information Gain, Chi-square

1. INTRODUCTION

Spam has become a headache for users on the Internet over the last few decades. Many solutions are developed and applied on problem, still spam continues to be major nuisance and we are still away from a satisfactory and long lasting solution. This is due to the fact that many heuristics are applied to proposed and developed methods and these heuristics are temporal and pertaining only to that particular corpus. Hence the devised solutions are not useful after a time period. Furthermore, spammers find new ways to easily overcome these solutions or devise new methods to send spams.

Most of the time, spammers use mail templates for sending spam. To send a particular promotion, they create pre-formatted template and merge the template with details of receivers stored in their database. Timely detection of these mails and underlying template features can be used to easily ignore forthcoming spam. Most high-volume spam is sent using such tools which randomizes parts of the message - subject, body, sender address etc. Templates of such mails can be included in the training set to minimize the search volume rather than using every mail in the corpora. The main objective of this paper is to investigate and evaluate the applicability of genetic algorithm and K-Means algorithm in the process of selection of suitable mail templates.

The remainder of the paper is structured as follows: Section 2 discuss abut the approach, Section 3 explains Genetic algorithm and K-Means algorithm and its applicability to this problem, Section 4 explains about the corpora used and how the mails are represented for the algorithm, experimental results are discussed in Section 5, and we conclude the work in Section 6.

2. APPROACH

The spam corpus contains both spam and legitimate mails. Our aim is to find out a small subset of these mails which best represent the corpus. Firstly, the attributes have to be analyzed using Information gain algorithm and Chi-square analysis and select important attributes from the corpus. Secondly, spam mails and legitimate mails are clustered separately using Genetic Algorithm and K-Means algorithm. Finally the experimental results are compared and choose the best method.

2.1. *Genetic Algorithm*

Genetic Algorithms (GA) apply an evolutionary approach to inductive learning. GA's were introduced as a computational analogy of adaptive systems. They are modeled loosely on the principles of the evolution via natural selection, employing a population of individuals that undergo selection in the presence of variation-inducing operators such as mutation and recombination (crossover). A fitness function is used to evaluate individuals, and reproductive success varies with fitness.

Genetic operations

Crossover: Crossover forms new elements for the population by combining parts of two elements currently in the population.

Mutation: Mutation is applied to elements chosen for elimination by randomly flipping bits within a single element.

Selection: Selection is to replace to-be-deleted elements by copies of elements that pass the fitness test with high scores. With selection, the overall fitness of the population is guaranteed to increase.

Fitness Function: Let N be the number of matches of the input attribute values of E with training instances from its own class. Let M be the number of input attribute value matches to all training instances from the competing classes. Add 1 to M. and divide N by M. The higher the fitness score, the smaller will be the error rate for the solution.

Supervised Genetic learning algorithm

Step 1: This step initializes a population P of elements. The P referred to population elements. The process modifies the elements of the population until a termination condition is satisfied, which might be all elements of the population meet some minimum criteria. An alternative is a fixed number of iterations of the learning process.

Step 2: First it applies a fitness function to evaluate each element currently in the population. During each iteration, elements not satisfying the fitness criteria are eliminated from the population. The final result of a supervised genetic learning session is a set of population elements that best represents the training data. Then it adds new elements to the population to replace eliminated elements if any. New elements are formed from previously deleted elements by applying crossover and mutation.

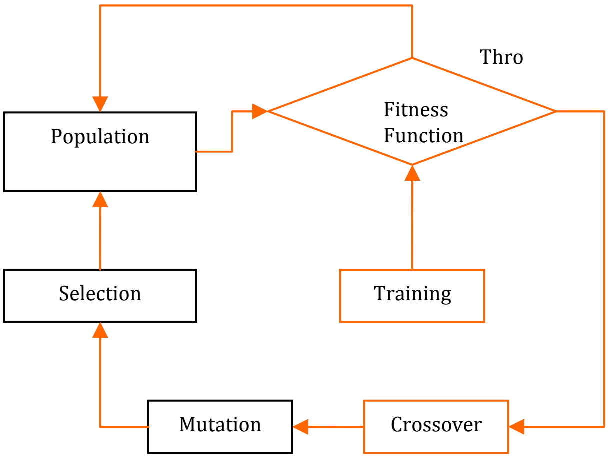

Fig 1: Supervised genetic algorithm

2.2 *K –Means Algorithm*

K-Means algorithm is used to solve the K-Means clustering problem [11] and works as follows. The method is used to find k clusters from the data set through an iterative procedure. The main idea is to define k centroids, one for each cluster. Then each point is compared with each centroid and the point is assigned to the cluster with nearest centroid. At this point, re-calculate k new centroids as centers of the clusters resulting from the previous step. The procedure is repeated until the centroids not moving to a new point or we can fix the number of iterations (cut-off). Finally, this algorithm aims at minimizing an OBJECTIVE FUNCTION, in this case a squared error function. The objective function $J=\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)}-c_j\right\|^2$, where $\left\|x_i^{(j)}-c_j\right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j, is an indicator of the distance of the N data points from their respective cluster centers.

3. Building Representation

We have considered emails as documents and terms as features. Each mail is tokenized (space as the delimiter) and stop words are removed. Term frequency and document frequency (tf and idf) measure is used to select initial feature vectors. The attributes with document frequency >= 10 and term frequency >= 4 are only selected. Then Information gain and chi-square analysis are applied on this vector to find out the most important attributes to form the feature vector. Both the methods gave almost same results. The Chromosome - Blueprint for a mail consists of reduced subset 26 attributes according to the information gain. The mails having attributes with highest gain are selected for the initial population. All the mails are encoded into their frequency representation.

For genetic algorithm, we need two sets of data, the initial population and the training set. The initial population is chosen from the training set according to some rules. GA’s are able to identify optimal or near optimal solutions under a wider range of selection pressure. However if the selection pressure is too low, the convergence rate will be very slow. Thus the initial population of mails was selected by the percentage of attribute contribution. The fitness score of each element in the population was computed and the elements with fitness score above the threshold (cut-off) value (say M =100 elements) were selected. For training set, entire training set is used.

For K-Means algorithm, we need training set and the value for K. Mostly K is chosen by applying heuristics which depends on the problem and in this case K is chosen as 100 for each class; spam and legitimate.

Composition of Initial population and Training set

Dataset	**No. of spam Mails**	**No. of ham mails**
Initial Population (for genetic algorithm)	50	50
Training set	200	200

Table 1: Data set Composition

The training dataset is prepared using the mails received in one month for the testing. Since the server is not capable to handle spam, we receive large numbers of spam mails every day. The initial population (N=50) is taken from this training set. Large datasets are available online, but when go for large datasets, the computational time increases and this will delay the mail delivery. Also template based spams are temporal, addressing current scenarios; so earlier templates may not helpful to detect spam.

4. Methodology

4. 1 *Genetic Algorithm*

Supervised genetic algorithm is applied on the initial population and training set to find out the best population by creating new generations. The fitness function used is Score =N / (M+1) where N is the number of matches of the input attribute values with training instances from its own class and M be the number of input attribute values matches to all training instances from the competing class. The instances with high scores are selected while low scored instances are eliminated. For eliminated instances, 50% Single entry cross-over and 28 bits mutation are done and which will be the second generation population. This process is continued until the population is converged.

4.2 *K –Means Algorithm*

K-Means algorithm is applied on the training set, separately on spam and legitimate mails. K is chosen as 50 for each class; spam and legitimate.

5. EXPERIMENTAL RESULTS

Before applying attribute selection and genetic algorithm, the Simple Naïve Bayes classification gave the following results:

Correctly Classified Instances	1583(97.2359 %)
Incorrectly Classified Instances	45(2.7641 %)
Kappa statistic	0.8797
Mean absolute error	0.0355
Root mean squared error	0.1507
Total Number of Instances	1628
Number of Attributes	310

Table 2 – Classification results before applying feature selection and learning algorithms

Confusion Matrix

spam	ham	<-... classified as
193	9	spam
36	1390	ham

Table 3 – Confusion matrix before applying feature selection and learning algorithms

The **Detailed Accuracy by Class** is as follows:

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.95	0.025	0.84	0.955	0.896	0.992	1
0.97	0.045	0.99	0.975	0.984	0.992	2
0.97	0.042	0.975	0.972	0.973	0.992	avg

Table 4 - Detailed Accuracy by Class before applying feature selection and learning algorithms

When Naïve Bayes Simple classification algorithm with 10-fold cross-validation is applied on whole dataset, the RMSE reported is 15%. We applied Information Gain algorithm to select high valued 50 attributes out of 310 attributes. The high valued attributes obtained by Information gain in their rank order are as follows:

> *your, cialis, software, attached, Viagra, cheap, soft, paliourg, file, xanax, meds, valium, tabs, prices, actuals, online, forwarded, quality, here, free, best, nomination, prescription ,...*

5. 1 *Genetic Algorithm*

An initial population of 99 mails is chosen for supervised genetic algorithm. (50 – legitimate mails and 49 spam mails). The performance of the online filtering strongly depends on the attributes and the training set selected. By applying supervised genetic algorithm, 100 best candidate instances from the large data set are selected for the final filtering of spam mails.

The results of classification using Simple Naïve Bayes algorithm with 10 fold cross-validation after the attribute reduction and initial population selection are:

Correctly Classified Instances	89 (89.899 %)
Incorrectly Classified Instances	10(10.101 %)
Kappa statistic	0.7984
Mean absolute error	0.1087
Root mean squared error	0.2971
Total Number of Instances	99
Number of Attributes	78

Table 5 – Classification results after applying feature selection and before genetic learning

Confusion Matrix

spam	ham	<-... classified as
49	0	spam
10	40	ham

Table 6 – Confusion Matrix after applying feature selection and before genetic learning

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.2	0.831	1	0.907	0.955	1
.8	0	1	0.8	0.889	0.95	2
0.8	0.09	0.916	0.89	0.898	0.95	avg

Table 7 – Detailed Accuracy by Class after applying feature selection and before genetic learning

After implementing Genetic algorithm, the Simple Naïve Bayes classification with 10 fold cross-validation produced the following results:

Correctly Classified Instances	94 (94.949%)
Incorrectly Classified Instances	5(5.0505 %)
Kappa statistic	0.8991
Mean absolute error	0.0648
Root mean squared error	0.2238
Total Number of Instances	99
Number of Attributes	78

Table 8 – Classification results after applying feature selection and genetic learning

Confusion Matrix

spam	ham	<-... classified as
49	0	spam
5	45	ham

Table 9 – Confusion Matrix after applying feature selection and genetic learning

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0.2	0.90	1	0.951	0.99	1
0.9	0	1	0.9	0.947	0.99	2
0.94	0.04	0.95	0.94	0.949	0.99	avg

Table 10 – Detailed Accuracy by Class after applying feature selection and genetic learning

5.2 *K –Means Algorithm*

K-Means algorithm is applied on the reduced attribute-set to find out the 100 centroids, 50 clusters each from spam and legitimate mails. R package is used for that. These cluster centroids-we say these are the templates- are used for classifying mails either as spam or legitimate. After executing K-Means algorithm, the Simple Naïve Bayes classification with 10 fold cross-validation applied on the centroids and it produced the following results:

Correctly Classified Instances	91 (91%)
Incorrectly Classified Instances	9(9 %)
Kappa statistic	0.82
Mean absolute error	0.1461
Root mean squared error	0.2818
Total Number of Instances	100
Number of Attributes	78

Table 11 – Classification results after applying feature selection and K-Means

Confusion Matrix

spam	ham	<-... classified as
49	1	spam
8	42	ham

Table 12 – Confusion Matrix after applying feature selection and K-Means

Detailed Accuracy by Class

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.98	0.16	0.86	0.98	0.916	**0.955**	1
0.84	0.02	0.977	0.84	0.903	**0.955**	2
0.91	0.09	0.918	0.91	0.91	0.955	

Table 13 – Detailed Accuracy by Class after applying feature selection and K-Means

6. Conclusions And Future Work

In this paper, we proposed template mail selection which uses supervised genetic algorithm and its operations, i.e., crossover and mutation, to create best templates in the training set for future spam filtering. The experiments show that proposed template mail selection performs efficiently and give better results. In addition, the system allows manual adjustments in the threshold value for fitness function, percentage of crossover operations, to the appropriate level of filtering. The overall quality of template mail instances is increased by applying genetic algorithm.

In K-Means algorithm the cluster centroids forms the training set for the final spam filtering task. This method can be applied and tested on big data to get an optimum training set.

References

[1] Usarat Sanpakdee, Aranya Walairacht and Somsak Walairacht, Adaptive Spam Mail Filtering Using Genetic Algorithm, Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference (Volume:1), DOI : 10.1109/ ICACT.2006.206004

[2] Mehmed Kantardzic(2003), Data Mining: Concept, Model, Medthod, and Algorithms, 'Genetic Algorithms', IEEE Press, 221-245 (2003)

[3] Data Mining: A Tutorial-based Primer" by Richard J Roiger and Michael W. Geatz, published by Person Education in 2003, pp.89-101

[4] Noraini Mohd Razali, John Geraghty, Genetic Algorithm Performance with Different Selection Strategies in Solving TSP, Proceedings of the World Congress on Engineering 2011 Vol II WCE 2011, July 6 - 8, 2011, London, U.K.

[5] Brad L.Miller David E. Goldberg, Genetic Algorithms, Tournament Selection, and the Effects of Noise, Complex Systems 9(1995)193-212

[6] L. Zhang, J. Zhu, and T. Yao, - An evaluation of statistical spam filtering techniques, ACM Transactions on Asian Language Information Processing [TALIP), v3, 2004, pp.243–269

[7] www.cs.cityu.edu.hk/~jfong/cs5483/Lectures/Lecture_10.ppt

[8] D. Nithya, V. Suganya , R. Saranya Irudaya Mary - Feature Selection using Integer and Binary coded Genetic Algorithm to improve the performance of SVM Classifier, Journal of Computer Applications (JCA) ISSN: 0974-1925, Volume VI, Issue 3, 2013

[9] Salehi, S.; Selamat, A; Bostanian, M., "Enhanced genetic algorithm for spam detection in email," *Software Engineering and Service Science (ICSESS), 2011 IEEE 2nd International Conference on* , vol., no., pp.594,597, 15-17 July 2011.

[10] Nadir Omer Fadl Elssied,Othman Ibrahim,Waheeb Abu-Ulbeh , An Improved Spam E-Mail Classification Mechanism Using K-Means Clustering, Journal of Theoretical and applied Information Technology, February 2014. Vol. 60 No.3

[11] http://www.onmyphd.com/?p=k-means.clustering visited on 3-11-2014

[12] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html visited on 3-11-2014