A Multi-objective Approach for Vietnamese Spam Detection

Minh Tuan Vu, Quang Anh Tran, Quang Minh Ha, and Lam Thu Bui

Abstract. In this paper, we propose a multi-objective approach for generating sets of feasible trade-off solutions for the Vietnamese anti-spam system (using SpamAssassin). The two objectives for considering are the Spam Detection Rate (SDR) and False Alarm Rate (FAR).The experiments were conducted based on Vietnamese spam data set through three scenarios with different numbers of SpamAssassin rules; and we used the non-dominated sorting genetic algorithm (version 2) – NSGA-II for finding the trade-off solutions. The result of each scenario was recorded to compare with the performance of the traditional approach (single objective optimization). According to the statistical results, the new approach not only achieved more efficient results but also created a set of ready-to-use rule scores which supports different levels of the trade-off between SDR and FAR.

1 Introduction

In recent years, when the spread of spams seems to be fierce and uncontrollable, researchers all around the world has managed to stop spammers from annoying email users by proposing a wide range of Anti-Spam solutions. For each solution with different approach, the pros and cons are various. There are also a number of factors to evaluate the efficiency of solutions. Among them, the Spam Detection Rate (SDR) and the False Alarm Rate (FAR) seems to be most obvious criteria to measure the effectiveness of a spam detection resolution.

The final purpose of any Anti-Spam approach is to maximize the SDR and to minimize the FAR as much as possible. The key point of problem is that the SDR

Minh Tuan Vu · Quang Anh Tran · Quang Minh Ha
Faculty of Information Technology, Hanoi University, Vietnam
e-mail: {minhtuan_fit,anhtq,minhhq_fit}@hanu.edu.vn

Lam Thu Bui
Le Quy Don Technical University, Vietnam
e-mail: lam.bui07@gmail.com

V.-N. Huynh et al. (eds.), *Knowledge and Systems Engineering, Volume 2*,
Advances in Intelligent Systems and Computing 245,
DOI: 10.1007/978-3-319-02821-7_20, © Springer International Publishing Switzerland 2014

is proportional to the FAR. Thus, the higher rate of detecting spam an approach brings the higher probability to alarm a ham (non-spam mail) as spam it gets and vice versa. An effective spam detection system is not expected to gain an absolute optimum which are 100% for SDR and 0% for FAR, but it is an acceptable trade-off between these criteria. Current approaches achieve the desired SDR (or FAR) by the following procedure:

1. A threshold at which an email is considered to be spam is predefined.
2. Model is built to train the system.
3. SDR (or FAR) is measured to evaluate the effectiveness of Anti-Spam solution at specific thresholds.

With this procedure, the only way to optimize the SDR and FAR without changing the model is to change the threshold. If email users' demand on the SDR and FAR are different, the threshold needs changing until matching their demands. For each time the threshold change, the whole training process is required to restart and consumes a lot of time.

In considering the concern of current Anti-Spam approach, the authors have applied the evolutionary multi-objective optimization algorithm –MOEA to solve the problem of SDR and FAR in Vietnamese spam detection. MOEAs have become popular as the solver for a number of multi-objective problems in different fields [1].By analyzing the nature of Anti-Spam problem and a wide range of MOEAs, authors figured out that NSGA-II [2] was suitable to build the framework and carry out the experiment. The performance of the algorithm was evaluated in [3] and said to outperform among other MOEAs.

The authors believe that the paper's contributions are two-fold. First of all, a set of Pareto is obtained. With this set of solutions, email users would have a list of SDR and FAR options for their different spam filtering demands. Each solution is available and ready-to-use without requiring retraining the dataset from the beginning. Secondly, Anti-Spam systems are provided a new approach to deal with the optimized tradeoff between SDR and FAR. The result of the paper illustrated that this approach was much more flexible and brought more satisfied results than single-objective optimization algorithms This paper is structured as follows: Section 2 introduced the background knowledge of the research. Section 3 explained the theoretical framework. Next, we presented the experiments and remarkable results in Section 5. Finally, the last section concluded the paper and talked about the future of our works.

2 Preliminaries

2.1 SpamAssassin Rules

SpamAssassin is one of the most popular mail filter developed by the Apache Software Foundation. It examines the message represented to it and assign a score to indicate the likelihood that the mail. SpamAssassin works basing on the predefined set of rules. A score is assigned to a rule. An email is marked as spam only when gaining enough the score which is greater than the threshold. Here is how a SpamAssassin looks like:

```
header FROM_STARTS_WITH_NUM From =~ /^\d\d/
describe FROM_STARTS_WITH_NUM From: starts with nums
score FROM_STARTS_WITH_NUM 0.390 1.574 1.044 0.579
```

The rule's name is FROM_START_WITH_NUMS. By applying the rule, SpamAssassin will examine whether the message's FROM header starts with at least two numbers against the regular expression. The score is added to the message's spam score if matching the rule. An anatomy of a rule was described in details by Schwartz (2004) [6].

2.2 NSGA-II

NSGA-II is an elitism algorithm introduced by Kaylyanmoy Deb in 2001[2]. The external set size (archive) equals to the initial population size. The current archive is determined based on the combination of the current population and the previous archive. The population is considered as a combination of several layers in such a way that the first layer is the best layer in the population by the dominance ranking.

The archive is formed against the order of ranking layers: Selecting the best ranking first. If the number of individuals in the archive is smaller than the size of population, the next layer will be taken into account and so on. A truncation operator is applied to that layer based on the crowding distance if adding a layer would increase the number of individuals in the archive to exceed the initial population size. Thus, the crowding distance of a solution x is the averaged total of objective-value differences between two adjacent solutions of the solution x, where the population is arranged according to each objective to find adjacent solutions and where also boundary solutions have infinite values. The truncation operator removes the individual with the smallest crowding distance.

An offspring population of the same size as the initial population is then created from the archive by using crowded tournament selection, crossover, and mutation operators. The crowded tournament selection rule is that the winner of two same-rank solutions is the one that has the greater crowding distance value [3].

3 Theoretical Framework

3.1 Problem

As mentioned in the introduction, the main concern of the traditional Anti-Spam approach is difficult and time-consuming to find out the optimized tradeoff between values of SDR and FAR if the threshold changes. If the set of spam detection rules remains unchanged, there is only one pair of values for SDR and FAR which are considered as the most wanted solution at a specific threshold. When the algorithm runs with different thresholds, the rule's scores (optimized for the predefined threshold) are no longer optimized for the current threshold which would cause the rate of spam detection and false alarm not optimized anymore. The training process must restart from the beginning to meet the email users' demand on various SDR and FAR.

3.2 Solution Design

This paper applied NSGA-II algorithm to solve the problem with two objectives: SDR and FAR. The first objectives SDR must be maximized while the second one FAR must be minimized.

Step 1: *Initialize the data input*
For the problem, the objective is also to find a set of ideal scores called x where

$$x = (x_1, .., x_m), \; m = (31, \; 51, \; 101), \; x_1 \; \in [2, \; 5], \; x_{2...m} \in [0,2].$$

The set of x will be generated randomly with a random algorithm which is a part of NSGA-II. Each value inside the set is considered as a chromosome. The first value is set limitation from 2 to 5 because it is the threshold – the point at what an email is considered as spam. The other values are set from 0 to 2 which are the score of SpamAssassin rules. Experiments were carried out with three cases (three different numbers of x): 30 rules and 1 threshold (m = 31), 50 rules and 1 threshold (m = 51), 100 rules and 1 threshold (m = 101).

Step 2: *Create the objective function*
The objective function is designed to run on the spam dataset S (231Vietnamese spam) and ham dataset H (251Vietnamese ham).

$$S = \{s_1, \; s_2, .., s_K\}$$

$$H = \{h_1, \; h_2, .., h_L\}$$

The set of N rules is pre-designed based on the framework in [4].

$$R = \{r_1, \; r_2, .., r_N\}$$

Each rule might match with some spams or hams through the matching function.

$$m(r,e) = \begin{cases} 1 \; if\_r\_matches\_e \\ 0 \; otherwise \end{cases} \tag{1}$$

Where $r \; \in R, \; e \; \in \{S,H\}$

The effectiveness of the set of rules with randomly-generated scores (from step 1) is evaluated by SpamAssassin against the dataset S and H. Score sets bringing the best results would be selected as a solution for this multi-objective problem.

At threshold T, the function to detect spam is implemented as follows:

```
//Input is an email
//Out is 1 if e is spam else 0
is_spam(e){
    score = 0;
    for i= 0 to N
    score += m(r,e)*score_of_r
    if(score > T)
    then return 1
    else return 0
}
```

Step 3: *Compute two objectives*
The purpose of the objective function is to compute two objectives of the problem. Within the scope of this problem, two objectives SDR and FAR are compute against the formula:

$$SDR = \frac{\sum_{i=1}^{K} is\_spam(s_i)}{K} \tag{2}$$

$$FAR = \frac{\sum_{i=1}^{L} is\_spam(h_i)}{L} \tag{3}$$

However, all objectives of NSGA-II algorithms are minimized [2]. Therefore, the SDR objective of this specific problem should be reformulate as (1 – SDR) to get the maximum.

Step 4: *Run NSGA-II algorithm*
After all data input and required parameters are ready, the NSGA-II program is called to run and figure out the best population. Based on that population, the final result would be evaluated and compared.

3.3 Algorithm Parameters

Due to the large number of parameters for the experiment, they were stored in a text file and passed into the program via the command line for each time the program called. The detailed descriptions of the parameters are shown in Table 1.

4 Experiments and Results

4.1 Experiment Settings

The experiments were carried out for three different numbers of rules' scores: 30, 50 and 100. Twenty simulation runs with twenty different random seeds are carried out for each set of rules. At the end of experiments for each set of rules, the results

Table 1 Algorithm Parameters

Algorthm Parameters	Values
Population size	100
Number of generations	1000
Number of objective functions	2
Number of constraints	0
Number of real variables	31 or 51 or 101
Lower limit of real variable 1	2
Upper limit of real variable 1	5
...	
Lower limit of real variable n	0
Upper limit of real variable n	2
Probability of crossover of real variable	0.9
Probability of mutation of real variable	1/number of real variables
Distribution index for crossover	5
Distribution index for mutation	10

were recorded for analyzed and compared to that of the traditional approach with single objective optimization.

Results gained from the experiments of this paper were compared to that from the experiment using the single objective optimization carried out in [5].

4.2 Experiments with 30 Rules

According to statistical results (Figure 1) from the experiments with 30 rules, in term of minimizing the FAR (at 0%), the best solution recorded for SDR was 62.34% for SDR while that result with single objective optimization (Table 2) is only 40.3% for SDR. Among solutions which the FAR are around 10%, the SDR of new approach with multi-objective algorithm NSGA-II are also much better the single one. They are {(74.03%, 7.79%); (74.46%, 8.66%); (72.29%, 6.93%)} in comparison to the best point {(67.1%, 9.6%)}. Further, the trade-off solutions found by NSGA-II were widely spread; this provides variety of good choices for the system.

4.3 Experiments with 50 Rules

According to statistical results (Figure 2) from the experiments with 50 rules, in term of minimizing the FAR (at 0%), the best solution recorded for SDR was 65.37% for SDR while that result with single objective optimization (Table 3) is only 43.7% for SDR. Among solutions which the FAR are around 10%, the SDR of new approach

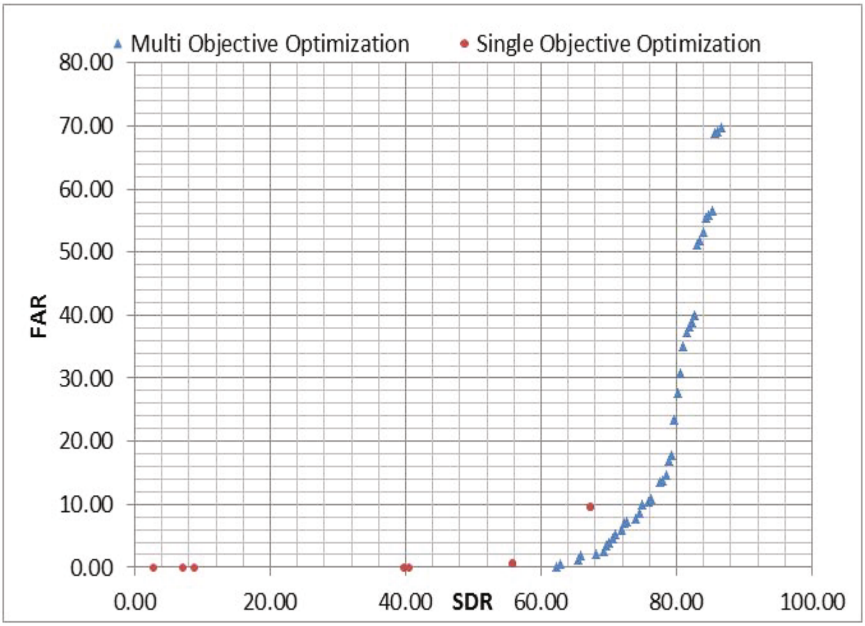

Fig. 1 The result of experiments using NSGA-II with 30 rules

Table 2 The result of experiments using single objective optimization with 30 rules

Threshold	Spam Detection Rate	False Alarm
0.5	67.1%	9.6%
1	67.1%	9.6%
1.5	55.8%	0.8%
2	55.8%	0.8%
2.5	40.3%	0.0%
3	39.8%	0.0%
3.5	8.7%	0.0%
4	6.9%	0.0%
4.5	2.6%	0.0%

with multi-objective algorithm NSGA-II are also much better the single one. They are $\{(83.98\%, 9.96\%); (83.55\%, 8.66\%); (82.68\%, 7.36\%)\}$ in compare to $\{(68.8\%, 9.6\%)\}$.

Although the result of single objective optimization had improved, they were still far from feasible solutions obtained by NSGA-II.

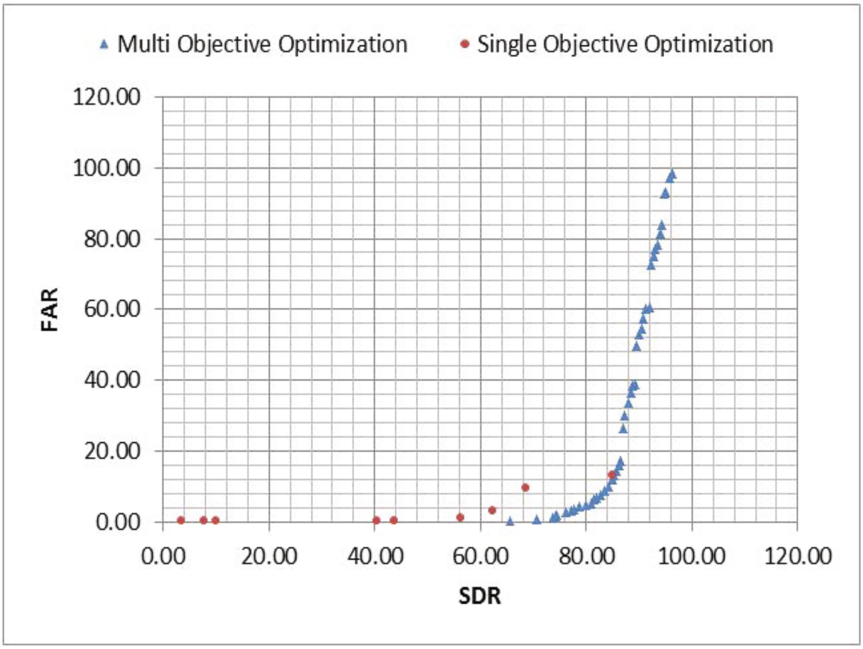

Fig. 2 The result of experiments using NSGA-II with 50 rules

Table 3 The result of experiments using single objective optimization with 50 rules

Threshold	Spam Detection Rate	False Alarm
0.5	84.8%	13.1%
1	68.8%	9.6%
1.5	62.3%	3.2%
2	56.3%	0.8%
2.5	43.7%	0.0%
3	40.3%	0.0%
3.5	10.0%	0.0%
4	7.8%	0.0%
4.5	3.5%	0.0%

4.4 Experiments with 100 Rules

According to statistical results (Figure 3) from the experiments with 100 rules, the best solution recorded for FAR was 0.87% with SDR at 64.5% while that result with single objective optimization (Table 4) was50.6% and 0% for SDR and FAR namely. In this scenario, although the new approach could not eliminate the rate of false alarm, the result, in term of maximizing the SDR, were even better than the one with 50 rules. They are {(83.55%, 8.23%); (81.39%, 6.06%); (82.25%, 6.93%)} in comparing to {(83.98%, 9.96%); (83.55%, 8.66%); (82.68%, 7.36%)} of NSGA-II with 50 rules and {(78.4%, 12%)}.

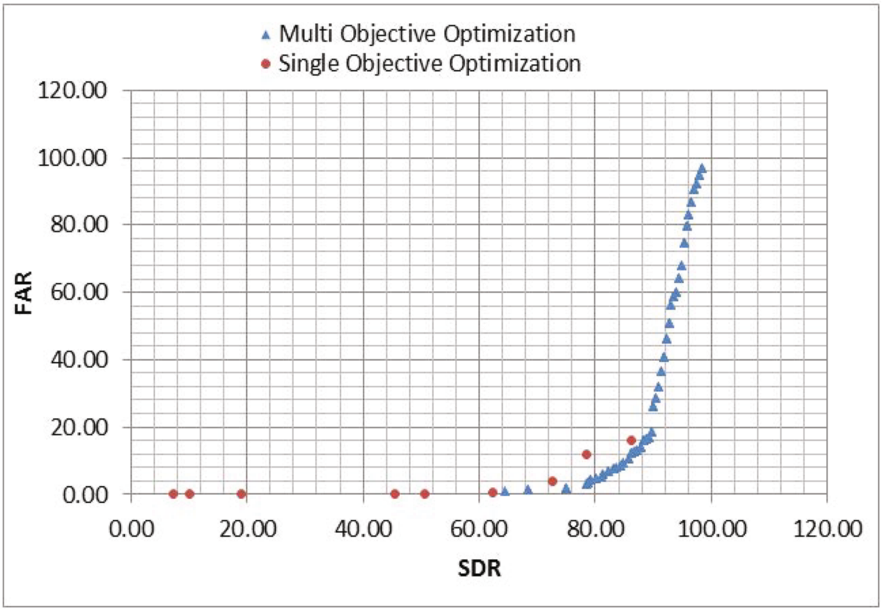

Fig. 3 The result of experiments using NSGA-II with 100 rules

Table 4 The result of experiments using single objective optimization with 100 rules

Threshold	Spam Detection Rate	False Alarm
0.5	86.1%	15.9%
1	78.4%	12.0%
1.5	72.7%	4.0%
2	62.3%	0.8%
2.5	50.6%	0.0%
3	45.5%	0.0%
3.5	19.0%	0.0%
4	10.0%	0.0%
4.5	7.4%	0.0%

Remarks

Based on the statistical results of the experiments, it is undeniable that the application of multi-objective optimization algorithm to spam detection is reasonable and promising. The new approach not only figured out more effective solutions for the issue of SDR and FAR but it also suggested a list of optimized options ready for choosing.

The illustration also pointed out that the more set of rules the algorithms working on, the better results it achieved. However, only the score of the rule changed for each time the algorithm run while the rule kept unchanged. Therefore, this method would save more time for training and updating new rules than the way the traditional approach did with single objective optimization algorithms.

5 Conclusion

In this paper, we proposed a framework which applied the multi-objective optimization algorithms – NSGA-II by Deb [2] to solve the problem of Vietnamese spam detection. In fact, traditional anti-spam approaches have optimized the spam detection rate and the false alarm rate for years and gained specific results. However, the achievement has been optimized for the single objective only. With the-multi objective optimization approach, not only one pair of SDR and FAR for each threshold has been worked out but a set of solutions with different tradeoff levels are computed. They all are feasible depending on specific email users' demands. More important, the score set of selected solutions are always ready to use without any training needed.

Despite of being a promising approach, the proposed framework remains some issues which need more efforts to resolve in the future. Firstly, it is the problem of runtime. Currently, there is no measurement about the runtime of the system. Because conducted experiments were carried out against quite small dataset, it is not a big issue. However, when the dataset expands in the future, this concern should be analyzed seriously. Secondly, the result of the experiment strictly depends on the performance of NSGA-II algorithm. The framework should be tested on other evolutionary multi-objective optimization algorithms for more diverse results.

Acknowledgement. This research was supported by the Vietnam National Foundation for Science and Technology Development (NAFOSTED) under project number 102.01-2012.04

References

1. Coello Coello, C.A., Veldhuizen, D.A.V., Lamont, G.B.: Evolutionary Algorithms for Solving Multi-Objective Problems. Kluwer Academic Publishers (2002)
2. Deb, K., Agrawal, S., Pratap, A., Meyarivan, T.: A Fast Elitist Non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II. In: Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J.J., Schwefel, H.-P. (eds.) PPSN 2000. LNCS, vol. 1917, pp. 849–858. Springer, Heidelberg (2000)

3. Bui, L.T., Essam, D., Abbass, H.A., Green, D.G.: Performance analysis of evolution multi-objective optimisation algorithms in noisy environments. Complexity International 11, 29–39 (2005)
4. Tran, Q.A., Duan, H., Li, X.: Real-time statistical rules for spam detection. IJCSNS International Journal of Computer Science and Network Security 6(2B), 178–184 (2006)
5. Vu, M.T., Tran, Q.A., Jiang, F., Tran, V.Q.: Multilingual rules for spam detection. In: Proceedings of the 7th International Conference on Broadband and Biomedical Communications (IB2COM 2012), Sydney, Australia, pp. 106–110 (2012)
6. Schwartz: SpamAssassin. O'Reilly (2004)