International Journal of Multimedia and Ubiquitous Engineering
Vol. 10, No. 3 (2015), pp. 119-126
http://dx.doi.org/10.14257/ijmue.2015.10.3.12

An Imbalanced Spam Mail Filtering Method

Zhiqiang Ma, Rui Yan, Donghong Yuan and Limin Liu

(College of Information Engineering, Inner Mongolia University of Technology, Hohhot, 010080, China)
mzq_bim@163.com,911736471@qq.com, 408236863@qq.com, liulimin789@126.com

Abstract

The high false alarm rate appears during the traditional spam recognition method processing the large-scale unbalanced data. A method which transforms the unbalanced issue into the balanced issue is proposed, when the K-means clustering algorithm is improved based on the support vector machine classification model, to obtain the balanced training set. Firstly, the improved K-means clustering algorithm clusters spam and extracts the typical spam，then the training set consists of the typical spam and legitimate messages, and finally the goal of the filtration of spam is realized by trained SVM classification model. Comparing the K-SVM filtration method to standard SVM method through the experiment, the result indicates that the K-SVM filtration method in large-scale unbalance data set can obtain high classified efficiency and the generalization performance.

Keywords: *Spam Filtering; K-means Clustering; Support Vector Machine; K-SVM Spam Filtering Method*

1. Introduction

E-mail is not only a widely channel for information exchange, but also an important way to obtain information. However, the email was implanted in commercial advertisements, malicious program or unhealthy information; system security and users' life have been seriously threatened and troubled. According to the China Internet association anti-spam Center released in 2013 thirty-second "2013Q1 Chinese anti-spam report" [1], the first quarter of 2013 Chinese internet users weekly received spam about 14.6 sealing, which accounted for the overall number of the ratio of 37.37%. How to distinguish spam from normal has become a major problem in the area of network security.

Spam filtering problem has become a classic and an important problem in the field of network security. The essence of the spam filtering is a text classification problem. In recent years, text classification technology has widely focused and developed, due to this phenomenon; a lot of text classification method has been successfully applied to spam recognition problems. At present, spam filtering model commonly used including several ways: (1) filtering model which is based on rules, this kind of method, is by looking for particular patterns in the mail content, including the header analysis, mass filter and keywords matching methods to spam filtering. The filtering method based on rules mainly includes Ripper method [2], decision tree method [3], the rough set method [4], the Boosting integrated method [5] and so on. (2) Spam filtering methods based on statistical is to solve the e-mail classification, and the classifier is automatically trained according to the sample set of spam and normal. The common filtering method based on statistics includes Bayesian classification algorithm [6] and Spam filtering method which is based on support vector machine (SVM) [7], *etc*.

At present, the studies mainly centralizes the balanced spam filtration problem under

ISSN: 1975-0080 IJMUE

statistical model, the number receiving the spam is larger than normal. The classification of spam then becomes an unbalanced classified question. If we use the traditional classification technology directly to process unbalance spam filtration question, the normal mail is easily recognized as the spam, which may lead to an inestimable loss to the user. How to deal with the problems of unbalanced mass spam is a difficult problem, and has an important theoretical significance and application value in the research on statistical models.

2. Spam Filtering Model based on Statistics

Statistics-based spam filtration process mainly includes pretreatment [8], feature selection [9], vector representation, the study training and filtration. Among them, the main purpose of pretreatment is to extract mail semantic characteristic [10], to realize the effective dimensionality reduction and to realize mail vector representation [11]. The study training is to get filtration model which is processed by the corresponded training algorithm in a labeled training set with extracting feature [12-13]. The filtration recognition is to recognize the vector representation mail by the filtration recognition algorithm, and make the category mark of mail [14]. The concrete process is as showed in Figure 1.

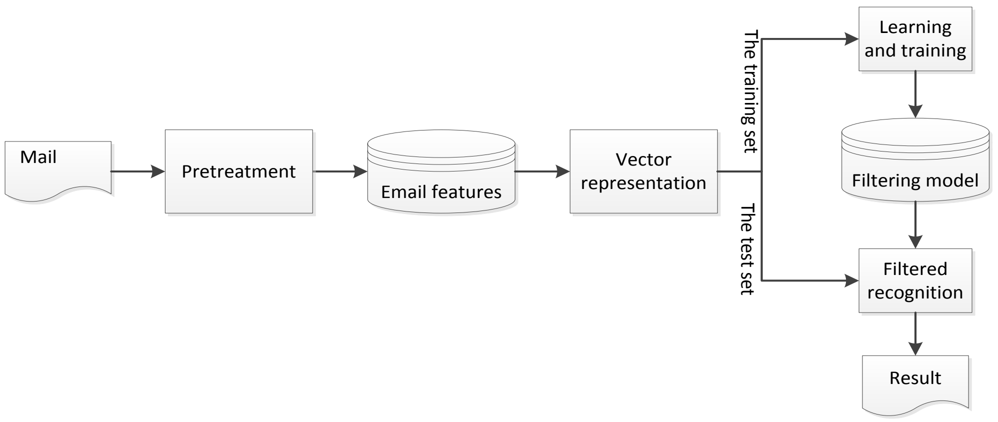

Figure 1. Spam Filtering Model based on Statistics

3. The Improvement of K-means Clustering Algorithm

Definition: the sample collection for X = $\{x_i\}^n_{i=1}$, x_i as samples and $x_i \in R^d$, $C_t \subseteq X$, t = 1, 2,... , k, $C_t = \{x_{ti}\}^m_{i=1}$, in the first t x_{ti} is the ith a sample; Function S ($x_{mi,}$ x_{mj}) is the first m class of the ith a sample with the first j a similarity distance between the samples.

If C_t, t = 1, 2,..., k is the result of clustering, clustering under C_t rigid satisfies:

(1)$\bigcup_{t=1}^{k} C_t = X$.

(2)C_z，$C_y \subseteq X$，$C_z \neq C_y$且$C_z \cap C_y = \emptyset$,

Obtaining

$$\mathrm{MAX}\left(S\left(x_{zi}，x_{zj}\right)\right) < MIN\left(S\left(x_{zh}，x_{yl}\right)\right)$$

The K-means clustering algorithm is a clustering algorithm which is commonly used. It is sensitive to the choice of initial cluster center, so it directly affects the efficiency of the algorithm execution. In the traditional K-means value cluster algorithm K initial cluster centers are selected with random, therefore, the method may possibly cause the cluster result to fall to the local minimum at the same time; while on the other hand may possibly cause algorithm the iteration step to increase, which may affect the efficiency of the algorithm implement.

In order to enhance the ability of K-means clustering algorithm processing the large-scale spam filtration, a proposed initial sample choice method is proposed. The main idea in the sample concentration is to take the physical distance between K points as the initial clustering centers, these centers may possibly belong to the different kind of the cluster results, so the initial cluster center itself has contained this effective prior information, accordingly reducing the iteration step of algorithm restraining, and enhancing the algorithm's efficiency. What's more, initial cluster center in the final cluster result belongs to the different category, and then, the different sample points with its similarity comparison can avoid the local minimum question that the traditional K-means clustering algorithm causes. The concrete flow is in the algorithm 1 as follows.

Algorithm 1: The improved K-means clustering algorithm

Step1: Choose the K initial cluster center. Let sample collection X= $\{x_i\}^{n}_{i=1}$, cluster integer K, the initialization cluster objective function $J^{(0)} = 0$, the initialization cluster center collection center=∅. The choice K cluster center concrete method is:

Step1.1: Solve sample collection X= $\{x_i\}^{n}_{i=1}$ central c, c^i was the ith characteristic of c, the computational method was as follows:

$$c^i = \frac{x_1^i + x_2^i + \cdots + x_n^i}{n} \tag{1}$$

Step1.2: Choose the distance center c which is the farthest sample x_i' from sample collection X= $\{x_i\}^{n}_{i=1}$, and joins the initial cluster center to gather center.

Step1.3: According to the formula (2), we can count Center set of samples with samples from the center set d (x_i, Center), these distances form a distance collections D(x_i, Center).

$$d(x_i, \text{center}) = \sum_{c_j \in \text{center}} \left\| x_i - c_j \right\|^2, \ \ x_i \notin \text{center} \tag{2}$$

Step1.4: Take non-center to concentrate d (x_i, center) which is the biggest x_i sample to join center.

Step1.5: If |center| ≥ K, return to a collection center as the initial cluster centers; or continue the Step1.3 execution until it is the size of the initial cluster centers center set.

Step2: Calculate each sample that x_i and the distance between the centers of each class c_j according to the formula (2), and take x_i as a member of the class which is most closely resemble class center.

Step3: Compute current objective function according to the formula (3), and calculate $r^{(t)}= J^{(t)}-J^{(t-1)}$, when $r^{(t)}$ is 0 or the limited value is the iteration step, then the cluster finished; Otherwise recomputed various class kind of central C_j according to the formula (4) , and returns to Step2 to continue to carry out.

$$J = \sum_{j=1}^{k} \sum_{i \in c_j} d_{ij}\left(x_i - c_j\right) \tag{3}$$

$$C_j = \sum_{q=1}^{n_j} {x_{jq}}/{n_j} \tag{4}$$

4. The K-SVM Non-equilibrium Spam Filtering Method

Traditional spam recognition methods cannot effectively deal with large-scale unbalanced spam filtering problem, therefore, we propose a k-means clustering algorithm based on support vector machine (K-SVM) spam filtering method, which is the combination of unsupervised K-means clustering method and support vector machine (SVM) classification method.

The core idea is to construct the training set from the normal mail and the typical spam determined by the improved K-means clustering algorithm, and then the SVM training is implemented. The detailed process is in the algorithm 2 as follows.

Algorithm 2: K-SVM spam filtering algorithm

Step1: Builds the training set X 'according to algorithm 3.

Step2: On the training set X ' train SVM learning, record the training time.

Step3: Construct the TX test set. Build a test set TX according to the Setp1 to Step4 in algorithm 3.

Step4: Take the SVM test on set TX tests, record the test accuracy.

Step5: Algorithm ends.

Algorithm 3: Building the training set algorithm

Step1: Construct the initial training regulations. Reading data centralizes n_1 normal mail gathers X^+ and n_2 spam gathers X^-, together form training regulations X, the mark of normal mail is +1, the mark of spam is -1, and spam's integer n_2 must be more than normal mail integer n_1 (i.e. $n_2 \gg n_1$), and calculates $\rho = \frac{n_2}{n_1}$.

Step2: Participle processing. The word is the semantic fundamental unit, Chinese is different from English; the words naturally divide through the blank space. In order to enhance the accuracy of cluster, the mail content needs to be carried on segment processing. We use the ICTCLAS participle system to carry on participle processing to the training regulations mail in the experiment.

Step3: Remove the Stop Word. After the segmentation, training regulations and test collection contain massive adverbs and auxiliaries, which also contain words that are divided by mistake. These words do not contain the important category message; we can remove them through the limit characteristic word of length.

Step4: Feature selection. Dictionary by pretreatment characteristics dimensions is still so high, and features of most e-mail dictionary dimension are between 5000-5500, using mutual information feature selection method for training set to reduce the objective dimension.

Step5: Spam cluster. For most spam data sets, X^- uses modified K-means clustering method for clustering, takes the cluster number $K=n_1$, clustering results are $X^-=\{x^-_i\}^k_{i=1}$.

Step6: Confirm the important spam. According to the formula (4) calculation of clustering results for each class in $c_i^-$the X_i^- centers, resulting in spam collection set of cluster centers $C^-=\{c_i^-\}^k_{i=1}$, and taking the center of each class as a new spam sample, we marked it as -1.

Step7: Construct a new training set. Combine the Step6 training sample c^- merged with the normal X^+ sample collection, and then forms a new set of training samples X'.

5. Experiment Result Analysis

In order to verify the performance of the K-SVM which is based on the unbalance spam classification approach to deal with the large-scale unbalance spam filtration issue, we compares standard the SVM method to the K-SVM method. Kernel function chooses Gaussian kernel, which is often used in the support vector machines.

The Gauss nuclear form is as follows: $K(x, y) = \exp\left(-\left\|x - y\right\|^2 / 2\,\sigma^2\right)$[15].The values of the kernel function of SVM and the penalty factor cannot be directly determined, so we selected the most appropriate value by setting a series of experiments. The polynomial and Gaussian kernel model experiment is respectively analyzed, obtaining Kernel parameter 1 and the penalty parameter 1000

The experiment uses six training data sets, which take 100 normal letters, and respectively 100 (1:1), 200 (1:2), 500 (1:5), 1000 (1:10), 2000 (1:20), 5000 (1:50) spam from the CCERT in 2006 May open e-mails. Proportion to spam and normal letters in the test set is fully compliant with the training set. Each training set has a corresponding test set randomly selected from the entire data set.

The experiment is conducted in a PC machine (2GHz CPU, the 1G memory), the experiment platform selects Matlab2008.

The evaluating indicator which the experiment uses not only includes training time that as mentioned before, also detects the following five important indexes:

(1) Normal Mail Rate (NMR), response the probability that the normal mail is classified as the normal mail. The definition is:

$$\mathrm{NMR} = \frac{\mathrm{N_{H\to H}}}{\mathrm{N_{H\to S}+N_{H\to H}}} \tag{5}$$

(2) False Alarm Rate (FAR), response the probability that the normal mail is classified as the spam. The definition is:

$$\mathrm{FAR} = \frac{\mathrm{N_{H\to S}}}{\mathrm{N_{H\to S}+N_{H\to H}}} \tag{6}$$

(3) Correct Filtration Rate (CFR), response the probability that the spam is classified as the spam. The definition is:

$$\mathrm{CFR} = \frac{\mathrm{N_{S\to S}}}{\mathrm{N_{S\to H}+N_{S\to S}}} \tag{7}$$

(4) Missing Rate (MR), response the probability that the spam is classified as the normal mail. The definition is:

$$\mathrm{MR} = \frac{\mathrm{N_{S\to H}}}{\mathrm{N_{S\to H}+N_{S\to S}}} \tag{8}$$

(5) Test Accuracy (TA), response the probability that all mails are classified as the right mails. The definition is:

$$\mathrm{TA} = \frac{\mathrm{N_{S\to S}+N_{H\to H}}}{\mathrm{N_{S\to H}+N_{S\to S}+N_{H\to S}+N_{H\to H}}} \tag{9}$$

And, $\mathrm{N_{H\to H}}$ expressed that normal mail judgment for normal mail number, $\mathrm{N_{H\to S}}$ expressed that normal mail judgment for spam number, $\mathrm{N_{S\to S}}$ expressed that spam judgment for spam number, $\mathrm{N_{S\to H}}$ expressed the spam judgment for the normal mail number.

Table 1 lists the experimental results from the K-SVM method and traditional standard SVM method in six groups of unbalance proportions training sets. The standard SVM spam classification method is expressed by SVM spam in the table, the k-SVM method is represented by k-SVM spam.

Table 1. The Experimental Results

Evaluating indicator	Algorithm	Spam proportion					
		1:1	1:2	1:5	1:10	1:20	1:50
TIME (s)	K-SVM spam	16.97	18.54	20.25	25.92	29.7	38.54
	SVM spam	15.41	20.36	39.78	63.05	291.74	983.3
NMR (%)	K-SVM spam	98	99	96	97	96	95
	SVM spam	98	96	90	83	68	41
FAR (%)	K-SVM spam	2	1	4	3	4	5
	SVM spam	2	4	10	17	32	59
CFR (%)	K-SVM spam	2	1.5	1.8	1.8	2.1	1.84
	SVM spam	2	2	2.4	1.3	0.45	0.2

MR (%)	K-SVM spam	98	98.5	98.2	98.2	97.9	98.12
	SVM spam	98	98	97.6	98.7	99.55	99.8
TA (%)	K-SVM spam	98	98.67	97.83	98.09	97.81	98.1
	SVM spam	98	97.33	96.33	97.27	98.05	99.18

Contrast 1: Training time. When training regulations sample size is tiny, for example when the total training mail sample integer is 200 or 300, two methods' running times are basic. But when the spam sample is increased, the training regulations scale also increases correspondingly, specially, when the type training regulations scale is over 5000, the SVM spam method's running time suddenly increases, the running time will be nearly 1000 seconds; K-SVM spam method running time along with training mail scale increases will not change hugely, and the running time still maintains within 50 seconds. With the increase of spam scale, the unbalanced state of training sample also obviously increases, therefore the spam data set's cluster integer also correspondingly increases, finally we only take the cluster center data to carry on the training to guarantee the SVM actual training regulations scale without increasing with the increase of primitive training regulations. Although it will take some time for the pretreatment process, the relative to the training process of SVM is negligible.

Contrast 2: Normal mail rate (NMR) and false alarm rate (FAR). When the training regulations scale is more balanced, for example when the rate is 1:1 or 1:2, normal mail from the two methods through rate and false alarm rate difference is small. But along with sample unbalanced increase, the normal mail rate that SVM spam obtains drops suddenly, simultaneously the false alarm rate also correspondingly rises. Specially, when the training sample unbalanced state is 1:50, the false alarm rate of training sample is over 50%, namely many normal mails were treated as the spam to be filtered by the system, so it will not be accepted by the users.

Contrast 3: Correct filtration rate (CFR) and Missing rate (MR). When the training regulations sample unbalanced state is small, two methods are almost consistent. But along with the training sample unbalanced enhancement, the Missing rate change of K-SVM spam method is not huge, but the SVM spam Missing rate gradually reduces, specially, when the sample unbalance performance achieves 1:50, the SVM spam Missing rate is only 0.2%.

Contrast 4: Test accuracy (TA), namely accuracy. When the unbalanced state of sample is small, two methods' accuracies are almost consistent. But gradually increases along with the spam sample, when the training sample unbalanced state increases gradually (when for example 1:5 and 1:10), the K-SVM spam accuracy outline is higher than the SVM spam accuracy, but along with training sample unbalanced further increase, the SVM spam accuracy can change is very high, specially, when the sample unbalanced state achieves 1:50, the SVM spam accuracy has achieved over 99%.

Contrasting the false alarm rate data, we may obtain the K-SVM spam method to have the obvious superiority compared to the SVM spam method. The contrast Missing rate and accuracy data, SVM spam may be better than the K-SVM spam method, but this is only one kind of false appearance. Because, when the training regulations sample size does not balance extremely, uses the sorter that SVM spam obtains most normal mail judgments is the spam, this lost the significance of spam filtration, and it has violated the fundamental principle of spam filtration system design, its Missing rate and whole correct precision “superiority” merely is because the spam scale is high, the proportion that accounts for is big, but the SVM spam sorter (including spam and non-spam) judges most mails for the spam creates, which cannot explain that this method is dealing with the unbalanced spam filtration issue the superiority, instead further confirmed its insufficient.

Therefore, K-SVM non-equilibrium spam filtering methods through effective compress the size of the spam, you can:

(1) Improve the speed of massive spam filtering.

(2) Reduce the non-equilibrium spam filtering false alarm rate of the problem.

6. Conclusion

This article studies the unbalance spam filtration question based the present situation to spam filtration. With the improvement of the K-means clustering algorithm, we proposed the K-SVM unbalance spam filtration method.

The pretreatment process consists of constituting the unbalanced training set and test set through the extraction of spam and normal mail, carrying on the segment, removing to Stop Word, extracting feature. The unbalance spam is classed by the improved K-means clustering algorithm based on support vector machine. The smaller training set comes from the more important spam and normal mail which are extracted from the clustering results. By training the support vector machines, we obtain the training model to distinguish the span from the normal. K-SVM method can make improvements in the large-scale spam extraction efficiency, and simultaneously guarantees the unbalance spam filtration efficiency.

We will study the extracting outstanding feature method to enhance the ability that the K-SVM spam model process the large-scale unbalanced spam.

Acknowledgements

Funding project: State Natural Sciences Foundation (61363052). Inner Mongolia Autonomous Region Natural Sciences Foundation project (2014MS0608). Inner Mongolia Autonomous Region college science research project (NJZY12052). Inner Mongolian university of Technology key Fund (ZD201118).

References

[1] China Anti-spam Condition Report [DB/OL]. http://www.12321.cn/pdf/2013Q1-anti-spam.pdf, **(2013)**.

[2] W. Cohen, “Learning rules that classify e-mail”, Proceedings of AAAI Spring Symposium of Machine Learning in Information Access, **(1996)**; Stanford, CA.

[3] X. Carreras and L. Marquez, “Boosting trees for anti-spam e-mail filtering”, Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing, **(2001)**, Tzigov Chark.

[4] D. Wang, W. Zhao, Y. Zhu, “Rough collection theory-based mail disaggregated model”, Computer engineering and application, vol. 42, no. 18, **(2006)**, pp. 167-170.

[5] T. Nicholas, “Using adaboost and decision stumps to identify spam e-mail. Stanford University”, **(2003)**.

[6] I. Androutsopoulos, J. Koutsias, KV. Chandrinos, “An evaluation of naive Bayesian anti-spam filtering”, Proceedings of the 11th European Conference on Machine Learning, **(2000)**; Barcelona, Spain: Springer-Verlag.

[7] H. Drucker, D. H. Wu, V. N. Vaonick, “Support vector machines for spam categorization”, IEEE Transactions on Neural Networks, vol.10, no. 5, **(1999)**, , pp. 1048-1054.

[8] J. MacQueen, “Some methods for classification and analysis of multivariate observations”, Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, **(1967)**; Berkeley.

[9] I. M. Rogat, Y. M. Yang, “High-performing feature selection for text categorization”, CIKM, **(2002)**.

[10] Y.M. Yang, J.O. Pederson, “A comparative study on feature selection in text categorization”, Proceedings of the 14th International Conference on Machine learning, **(1997)**, Nashville, Tennessee, USA.

[11] T. Shi, “Support vector machines' applied research in email classification”, Computer simulation, vol. 28, no. 8, **(2011)**, pp. 156-159.

[12] W. Ruilei, J. Luan and X. Pan, “One kind of Chinese participle of improvement to biggest matching algorithm”, Computer application and software, vol. 28, no. 3, **(2011)**, pp. 195-197.

[13] J. Lei and X. Sun, “One kind of intelligent spam filtration model simulation”, Computer simulation, vol. 30, no. 5, **(2013)**, pp. 370-374.

[14] Y. Shao, "The research and simulation of spam optimization filtration method", Computer simulation, vol. 30, no. 12, **(2013)**, pp. 265-268.

[15] Y. Lin, "The model selection of support vector machines studies", Harbin Industry University: Harbin Industry University, **(2006)**.

Authors

Zhiqiang Ma (1972-), male (HUI), Inner Mongolia Hohhot associate professor, master's tutor, main research field is the search engine, speech recognition, cloud computing.

Rui Yan (1988-), male (the Han nationality), Inner Mongolia Erdos people, graduate student, and main research field is the speech recognition, data mining.

Donghong Yuan (1985-), female (the Han nationality), Shanxi Shuozhou people, graduate student, and main research field is the spam filtering, data mining.

Limin Liu (1964-), male (the Han nationality), Inner Mongolia Hohhot, Professor, master's tutor, main research field is the electronic commerce, data mining.