IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 1, January 2017

Incremental Learning for Spam Detection

Amin Shams[1], Touraj Banirostam[2]

Student, Department of Computer Engineering, Electronic Branch, Islamic Azad University, Tehran, Iran[1]

Assistant Professor, Dept of Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran[2]

Abstract: Through the considerable growth of using email in recent years, the large volume of unsolicited emails called spam made the researchers inspired by text classification techniques to implement systems for filtering such junk emails. Therefore, the goal of current research is to survey all kinds of spams and the issues they cause. It tries to present a method for optimizing spam detection and resolving complexities and diagnostic problems, as well as raising the sentiment of speed and accuracy in spam detection using incremental approach based on collective learning. So this method is to be able to automatically identify spams with mentioned approach. This is because in contrast with batch learning, data is defined as a batch or a group of data at any time which rises the accuracy of spam detection in incremental learning approach. So the goal of algorithm in proposed method is to produce an incremental training model similar to the trained model in batch state. Results of assessments indicate that proposed algorithm can show higher efficiency like other investigated incremental algorithms. Also experiments indicate that proposed algorithm is successful in detecting new input samples through introducing new classes.

Keywords: Spams, Spam Detection, Incremental Learning, Collective Learning.

I. INTRODUCTION

All around the world, increasing use of electronic letters due to their facility and low cost makes many internet users become interested in developing their business within the context of internet. In this environment, many companies have been drawn to the idea of advertising through electronic mail which makes internet users' mailboxes get full of unsolicited emails called spam. Many users face emails wasting their time to be organized and also wasting their band space and mailbox space too. This was the starting point to develop automated spam management approaches. Emails classification based on documents classification seems to be a suitable technique to solve such problems [1]. These spams only cause some inconvenience in the past while today, they become a big problem because of their massive volume which waste users' time and money. So methods are needed to filter these unsolicited emails automatically. Spam filtering in fact is a computer program able to classify emails which is most likely able to detect spams. Most of these filters utilize a combination of several methods such as black list or white list, using keywords, law-based filtering, and etc. to have a more accurate spam identification. These methods can be solely efficient filters but for commercial functions, their combination is more used. Some of them would be manually specified by user like the filtering used in Yahoo email. But there is a major fault in such methods which is using fixed rules that should be costumed by user instead. Their other problem is that the spammers can fool them by different tricks [2].

There is another method in addition to mentioned ones which becomes very popular and works according to identify spams based on their content. This content-based method have been considerably advancing in recent years. Separating legitimate emails and spams based on content can be considered as a kind of text classification since the body of most emails is a text whose class should be determined while a spam enters.

Machine learning methods have many applications in texts classification and achieve good results in this area. Therefore, machine learning algorithms can be applied in spam classification so that in recent years, machine learning algorithms have been widely used in electronic spams filtering. These methods have largely improved spam filtering. These algorithms learn to classify documents based on their content. This kind of action needs learning phases. Basically, learning can happen in two major methods: batch and incremental. Learning in many of the algorithms used in spam filtering, happens in batches so that in case of addition of a new training data, learner is not able to add new knowledge to its prior set of knowledge and if it wants to utilize new knowledge as well, it has to do its leaning again on whole old and new data; while problems such as lack of access to the whole training data at the moment, limitations of using learner memory and etc., make these methods useless in most of applications in real world like spams. That is the reason of the tendency to use algorithms based on incremental learning for spam classification. Using these algorithms allows learner to learn knowledge and add it to its prior set of knowledge while facing new training data without forgetting its past knowledge. There are several incremental learning based algorithms such as Support Vector Machine (SVM), Bayes, and Decision Tree.

Currently, one of the biggest challenges which computer engineers are facing is investigating and innovating methods which in separating and identifying important emails from spams can happen with high speed and accuracy. In order to this goal, many researched have been

DOI 10.17148/IJARCCE.2017.6101

done in different methods introducing different algorithms considering that incremental learning is one of methods with no long history in spam detection discussion and have been less studied. So in current research, it is tried to rise important information and email detection's accuracy and speed and separating spams as well through using incremental learning.

II. HISTORY OF RESEARCH

Kiani Galougahi and Shiri Gheidar [3] have worked on a paper called"Controlling concept change in spam detection using sequential clustering under supervision". In this paper, a new method to control concept change in spam detection using sequential clustering under supervision has been presented. Concept change means changes in concept of the goal because of change in symbolic definitions of goal concepts, generating a new goal concept, and changing goal concept of samples. Using clustering and a new similarity criteria in this method, training samples are divided into separated subsets. After that, convenient features are extracted from representative vectors of each subset through using genetic algorithm and a new fitness function.

Khosrotaj [4] studied "Dynamic spam filtering based on ontology" in his paper. In this research, a domain ontology is used to specify the class type of emails and consequently classifying them. Furthermore, while doing this job through using set of data, the basic ontology becomes completed and ultimately a new ontology is used for email classification. It should be considered that change of people's taste through time is in a direct relationship with spam filtering. This relationship is so that ontology should update itself and generate a dynamic mechanism to stop sending spams due to change of people's taste.

In his research called "presenting a new method for spam filtering", Hashemi [5] cites that a new method for internet spam filtering enables users to use several services. Sending and receiving emails is one of the oldest and at the same time, the most common service presented on internet. In spite of all aforementioned advantages and service potentials in recent years along with the growth and development of internet use, some problems have been observed in this respect such as distributing emails infected with viruses or worms, and sending/receiving unsolicited emails so-called Spam. Spam, due to experts, internet practitioners, and professionals in this area is more an ethical issue.

Heidari [6] discussed "intelligent spam filtering system using machine learning algorithms" in his paper. The goal of his research is to study existing methods of spam filtering as well as developing and implementing a new efficient method of spam filtering based on combined methods of machine learning. This research has used data set of LingSpam and SpamAssassin including three parts of data preparation, single classification, and combined classification. Among the results of single batching, SVM, SMO, and Logistic Regression algorithms have presented the best performance. Finally with investigating the results of combined classification, propose model has been presented according to accuracy, precision and recall.

Munde Kusum et al. [7] have presented "An Incremental Spam Filter" which in a new algorithm has been introduced based on incremental learning that provides the best performance. This algorithm adds the new knowledge according to new training data of weighted majority voting.

Li et al. [8] investigated "AN INCREMENTAL LEARNING BASED FRAMEWORK FOR IMAGE SPAM FILTERING". Time cluster has been used in this paper to solve the problem of spam detection and afterwards, a combination of clustering and machine learning has been presented for enhancing capacity. The mechanism of integrated incremental learning is to ensure the ability of proposed framework to organize itself to overcome image spam fraud.

Yang and Jungchen [9] presented "Uninterrupted Approaches for Spam Detection Based on SVM and AIS" which in eight methods have been studied and compared due to their speed and accuracy in spam detection.

Alqatawna et al. [10] presented an incremental learning algorithm based on group learning. Then a software of proposed algorithm for spam filtering has been discussed. This algorithm is called incremental root strengthening and it assumes the environment to be constant.

III. RELATED WORK

Considering library studies along with aforementioned matters, it has been tried in this paper to present a method of optimizing spam detection using incremental learning approach to resolve complexities and diagnostic problems as well as rising speed and accuracy sentiment in spam detection. This method is trying to be automatically able to detect spams with the incremental approach based on collective learning. In contrast with batch learning, data is defined in batches or groups at any time in incremental learning which itself rises the accuracy in spam detection. So the aim of algorithm in proposed method is to make an incremental training model similar and near to trained model in batch state. In this method in another word, learning happens through ongoing update of a model based on different subsets of training data. This learning would be used once training data and its calculation cost are huge or when a new set of training data would emerge after the end of training time. A computer system equipped with updated hardware has been used to run these algorithms. Also to simulate and implement the proposed algorithms, MATLAB and other similar software have been used. Moreover, all presented diagrams will be designed with EXCEL software.

A. Collective Learning

There are many algorithms for incremental learning and among these methods, algorithms based on collective learning produce more accurate and robust results which would be because of utilizing several classifier [11].

Furthermore, it has been shown in most of problems that a collective learner works more accurate than any basic classifier [12]. So according to cited matters, the proposed algorithm presented for spam classification is to be an incremental one based on collective learning. It can be said that in this algorithm, several weak classifier has been generated for every new set of data which their results are to be combined with the results of previous classifiers with a method like weighted majority voting. Collective learning used in this algorithm has a robust reinforcement structure so that each basic classifier is trained on differently distributed training samples. A homogeneous structure has been also picked in classifiers' design too which means that all basic classifiers have been trained with the same learning algorithm.

B. Incremental Learning Problem

Title In the beginning of this section, the incremental learning problem in spam detection has been expressed to continually introduce the proposed algorithm to solve it. Also some points should be basically noted: first, considered learning in this problem is batch incremental learning; second, discussing considered environment in this algorithm which has been assumed to be constant and changeless; and the last one is that the proposed algorithm has not been considered cost-sensitive.

Incremental learning in spam detection is considered a text classification issue since once a letter (which is usually a text in natural language) enters, the class (spam or legitimated letter) which it belongs to should be determined. So each letter is assumed as X which is a vector of features in R^m space and determined by y label (a set of limited values). In spam detection problem, this label is defined as $\{-1,+1\}$ that +1 label shows the letter to be spam and -1 indicates that to be legitimated.

In contrast with batch learning, data is being defined at any time in batches or groups in incremental learning [12]. For instance, in t time, used data is considered as S_t. the goal of the method used in this research and most of incremental algorithms is to produce an accurate classifier able to determine true label of an unknown letter through utilizing existing new information in data set of S_t without forgetting achieved information of previous data that are $S_1,\ldots S_{t-1}$.

C. Proposed Model

A new incremental algorithm for spam detection has been introduced in this section. This method tries to be able to automatically identify spams with an incremental method based on collective learning. The aim of this algorithm is to produce an incremental training model similar and near trained model in batch state. In another word, learning in this method happens through a model based on constantly updated different subsets of training data. This learning would be used once training data and its calculation cost are huge or when a new set of training data would emerge after the end of training time. Proposed model produces a set of several classifiers for each set of data which means that a collective learner is produced every set of data. This collective learner has been considered as a batch algorithm called RotBoost [11]. So it can be said that the proposed model is a set of several collective learners combined together. For a better understanding, first RotBoost algorithms has been reviewed.

As shown in the pseudo code of this algorithm in figure 1, this combination is such that apparently AdaBoost algorithms seems to be used as the weak learner within Rotation Forest algorithm.

FIG 1 PSEUDO CODE OF ROTBOOST ALGORITHM

```
Training Phase
For s= 1,2,…,S
• Use the steps similar to those in Rotation Forest to computer the
 rotation matrix, say $R_S^a$ and let $\varphi^a = [XR_S^a Y]$ to the training set for
 classifier $C_S$.
• Use the steps similar to those in AdaBoost to
generate $C_t^a$ and $\alpha_t$
End for
• Let $C_S(X) = \arg\max_{y\in Y} \sum_{t=1}^{T} \alpha_t I(C_t^a(x) = y)$
End for
Output
The class label for x predicated by the final ensemble
hypothesis
$C^*(X) = \arg\max_{y\in Y} \sum_{s=1}^{S} I(C_S(x) = y)$
```

IV. RESULTS OF SIMULATION

In this section, the performance of proposed algorithm in spam classification has been assessed and this algorithm has been presented with two types of incremental learning algorithms such as Bayesian.

- **Ling Spam set of data**

Results of running the proposed algorithm, Bayes, and Learn++ on Ling Spam set of data have been shown in table 1.

Table. 1. Comparing the proposed model with two incremental algorithms on Ling Spamset of data

Algorithm/ Criteria	The proposed Model	Bayesian Algorithm	Learn++ Algorithm
Accuracy	95.6	94.12	78.64
Error	4.4	5.88	21.36
False Positive Rate	0.048	0.032	0.26

Spam Accuracy	0.85	0.84	0.55
Weighted Accuracy	95.21	96.74	74.36
Weighted Error	4.49	3.26	25.61

The important point at the first look at table 1 is weak results of Learn++ algorithm in comparison with two other ones. So superiority of the proposed model is obvious. Also comparing to Bayesian algorithm, it can be seen that although accuracy has been slightly increased, but false positive rate has been a little decreased. In addition, weighted accuracy has been increased in Bayesian algorithm while this value has not changed much in incremental algorithm and has only decreased about 0.04% which has been very small. Fig 2 has been presented for a better understanding. So looking at this Figure 2, it can be cited that the proposed algorithm has achieved a desirable efficiency on Ling Spam set of data in comparison with two other incremental algorithms.

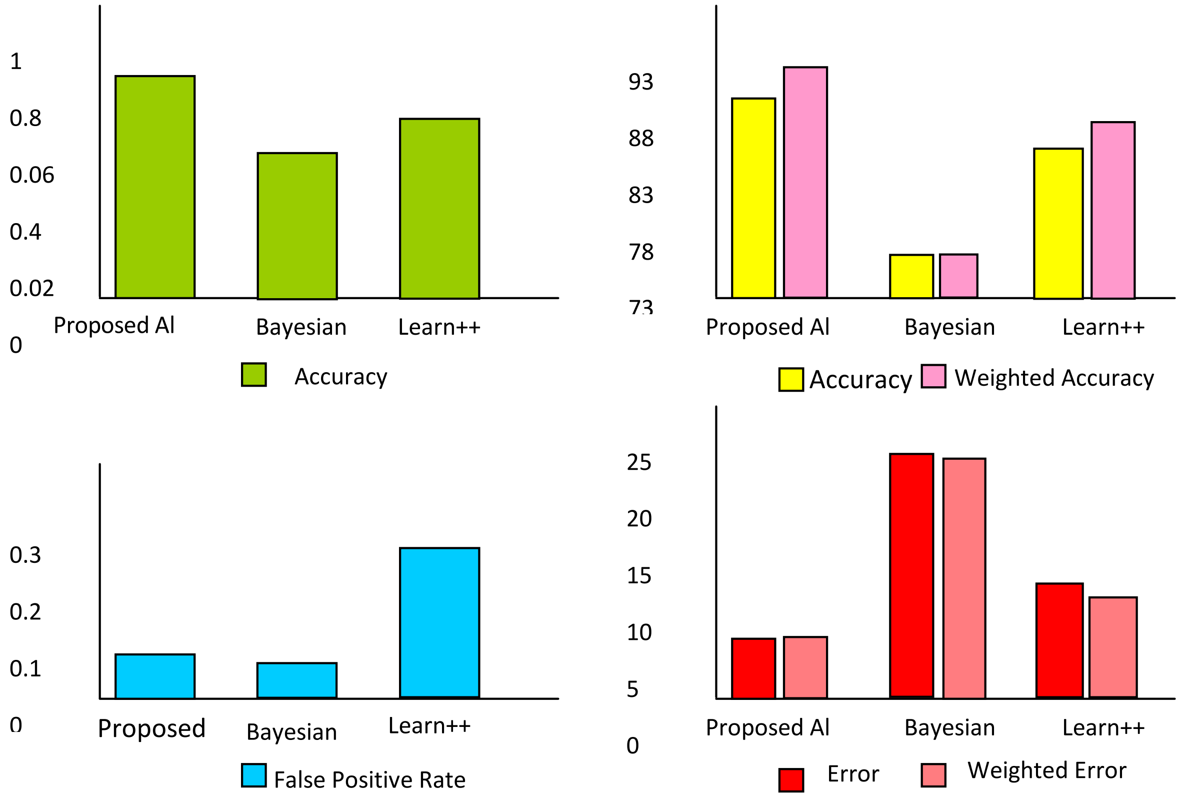

Fig. 1. Comparing the proposed algorithm on Ling Spam set of data

- **Spam Assassin set of data**

Now comparisons above should be done on the new set of data. So in this section, the comparison is done on Spam Assassin set of data and results have been indicated in table 2. As shown there, Bayesian algorithm achieved worse results comparing to the two other algorithms in contrast with its results on Ling Spam set of data so that as of now, the proposed model has become superior to Bayesian. Also according to Fig3, the difference between the proposed model's behavior and Learn++' is very small which is again implicating on the superiority of the proposed model.

So the common point in behavior of this set and Ling Spam set of data has been the same performance of Bayesian algorithm in increasing weighted accuracy and also the performance of Learn++ algorithm in decreasing weighted accuracy, while performance of the proposed model has not been considerably increased and was only decreased about 0.02% which is a very little amount. Therefore, comparing to Learn++ and Bayesian algorithms, this is the proposed algorithm that could again reach the highest score just like two previous sets of data. Since behavior of Learn++ and proposed algorithms are very close to each other, the table of mentioned algorithm' states is used for a better understanding. So the average states of running these three algorithms for Spam Assassin has been shown in table 3.

This table can properly show better efficiency of the proposed model. Reviewing these results, it can be seen that the results of Bayesian algorithm are worse than two other algorithm' while results of the proposed algorithm and Learn++ are very close. The only difference is in false negative rate so that the proposed algorithm decreases this rate to zero while it is not zero for Learn++ algorithm.

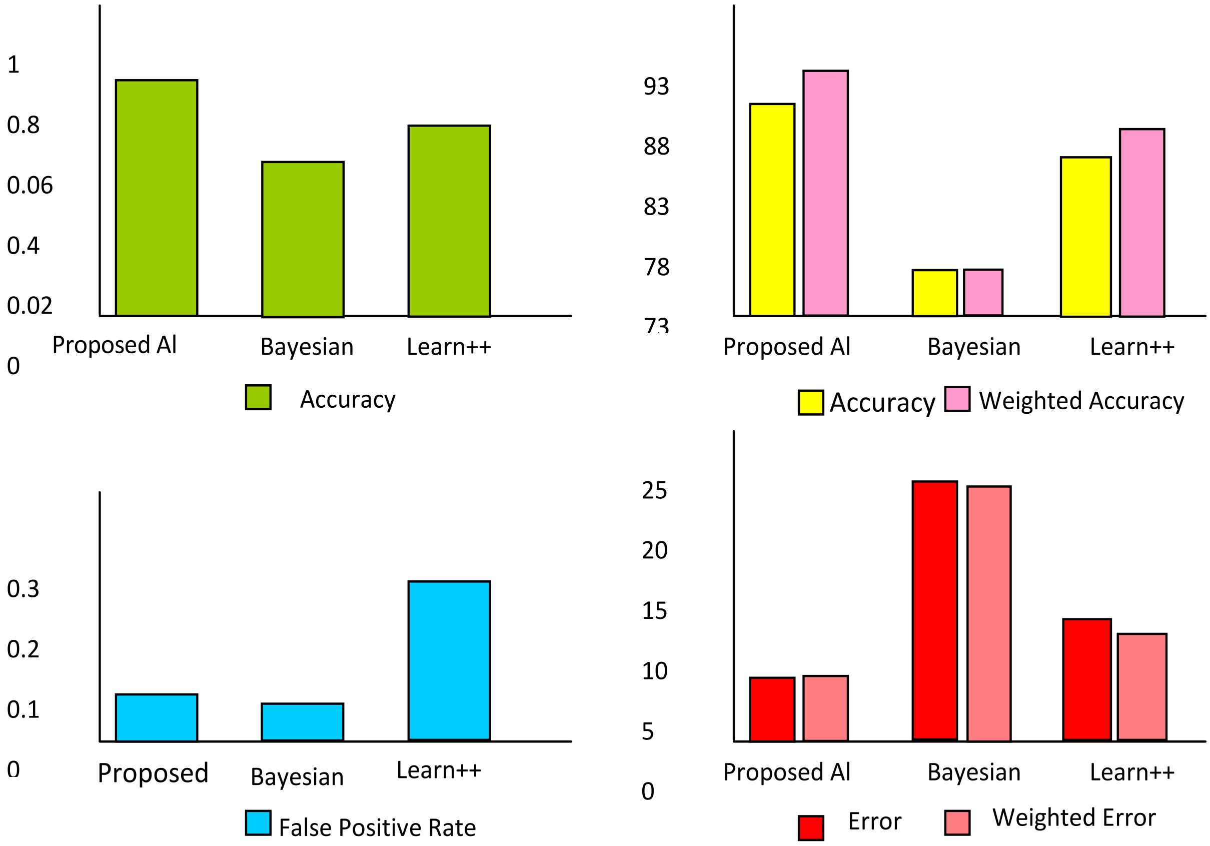

Fig. 2. Comparing the proposed model on Spam Assassin set of data

Table. 3. Table for three studied algorithms in Spam Assassin set of data

The proposed model	Spam	Legitimated letter	Real Predicted
	0 949	2072 3	Legitimated letter Spam
Bayesian Algorithm	Spam	Legitimated letter	Real Predicted
	95 854	2021 54	Legitimated letter Spam
Learn++ Algorithm	Spam	Legitimated letter	Real Predicted
	1 948	2072 3	Legitimated letter Spam

V. CONCLUSION

As cited above, nowadays spams have become a big problem for email users. So it has been many years that efforts have been done to overcome this problem in virtual world. The result of these efforts has proved that utilizing machine learning algorithms in identifying and filtering spams have been more successful.

Considering achieved results in this research, incremental algorithms based on collective learning has been proposed for spam classification. This algorithm performs collectively through producing a collective learner for every set of input data. To examine the incremental ability for spam classification, this algorithm has been assessed with two Ling Spam and Spam Assassin sets of data. But this should be mentioned that classifying sets of data has been based on body and the subject of existing letters in each set of data. Achieved results of this assessment confirms that the proposed algorithm could show higher efficiency like the other investigated incremental algorithms.

REFERENCES

[1] Kharazmi, Sadegh; Farahmand, Ali; Behjati, Shahab (1393), Spam emails detection using Bayesian spanning tree, National Conference of Computer Science.

[2] Kiamarzpour, Forouzan; Bayani, Nayat-allah; Habibi Dahaneh Siri, Zeinab, (1392), Investigating the effects of features' size by IG and Relief Algorithms on email classification using decision trees, National Conference of Computer Science and IT.

[3] Kiani Galougahi, Hamed; Shiri Gheidari, Saeed, (1386), Controlling concept change in spam detection using sequential clustering under supervision, The 13th National conference of Iran's computer association, Kish Island, Iran.

[4] Khosrotaj, Roya, (1392), Dynamic spam filtering based on ontology, Master Thesis, Chamran University of Ahvaz.

[5] Hashemi, Sara, (1392), Presenting a new method for Spam filtering, Master Thesis, Shiraz University, research center of Electronic and Computer.

[6] Heidari, Fatemeh, (1393), The intelligent Spam filtering system using machine learning algorithm, Master Thesis, Non-governmental non-profit institution of higher education of Ahvaz university.

[7] [Munde Kusum M. Mangrule Rupali A. (2011), An Incremental Spam Filter, International Journal of Innovations in Engineering and Technology (IJIET).

[8] Li Xiao Mang, HaRim Jung, Hee Yong Youn and Ung-Mo Kim, (2014), AN INCREMENTAL LEARNING BASED FRAMEWORK FOR IMAGE SPAM FILTERING, International Journal of Computer Science, Engineering and Applications (IJCSEA) Vol.4, No.1.

[9] Ying Tan, and Guangchen Ruan, (2014), Uninterrupted Approaches for Spam DetectionBased on SVM and AIS, Columbia International Publishing, International Journal of Computational Intelligence and Pattern Recognition, Vol. 1 No. 1 pp. 1-26.

[10] Alqatawna, Ja'far, Hossam Faris, Khalid Jaradat, Malek Al-Zewairi, Omar Adwan, (2015), Improving Knowledge Based Spam Detection Methods: The Effect of Malicious Related Features in Imbalance Data Distribution, Int. J. Communications, Network and System Sciences, 8, 118-129.

[11] Zhang, C. X. Zhang, J. S. (2008), RotBoost: A Technique for Combining Rotation Forest and AdaBoost", Pattern Recognition Letters, vol. 29, pp. 1524–1536.

[12] Liu, R. and Yuan, B. (2001), Multiple Classifiers Combination by Clustering andSelection, Information Fusion, vol. 2, pp. 163-168.

DOI 10.17148/IJARCCE.2017.6101