An Incremental Spam Filter

Munde Kusum M.
Department of Computer Science and Engineering
Marathawada Institute Of Technology, Aurangabad, Maharashtra, India

Asst. Prof. Mangrule Rupali A.
Department of Computer Science and Engineering
Marathawada Institute Of Technology, Aurangabad, Maharashtra, India

Abstract- Nowadays, email plays a significant role in communication and it is being preferred by everyone. As the spam problem grew larger, the interest in spam filters grew accordingly. Different techniques are present to deal with spam filtering. In this paper, a new algorithm based on incremental learning is introduced which provides best performance. The algorithm acquires new knowledge from new training data and adds it to previous knowledge based on weighted majority voting. It generates additional ensemble when new data become available without losing old acquired knowledge. The proposed algorithm works efficiently compared to other related algorithms.

Keywords – Spam, Spam Filter, Boosting, Ensemble Learning, Machine Learning, Incremental Learning

I. INTRODUCTION

Emails are usually classified into ham and spam mails. Ham mails are legitimate emails we want to receive whereas spam mails are unsolicited mails or unwanted mails. The rapid growth of email usage in recent years has brought us the large amount of spam mails containing bad contents, malicious code like virus that could cause various kinds of damages to systems.

Spam is also known as junk email or unsolicited bulk email, clicking on links in spam email may send users to phishing web sites or sites that are hosting malware. Spam email includes malware as scripts or executable file attachments. Spammers collect email addresses from chat rooms, websites, mailing lists and newsgroups.

A. Machine Learning -

Spam filtering is a classification problem, which can be solved using different machine learning algorithms. Machine learning is a branch of artificial intelligence, concerns the construction and study of systems that can learn from data [8]. There are various machine learning algorithms like Naive Bayes, Decision Tree, SVM etc. For classification we need a set of predefined classes and want to know which class a new object belongs to. In the context of machine learning, classification is supervised learning. A supervised learning system that performs classification is known as a learner or classifier.

B. Ensemble learning -

It is a method which learns multiple alternative definitions of a concept using different training data or different learning algorithms then combine decisions of multiple definitions using weighted voting. There are two types of ensemble learning, Homogeneous ensemble learning and heterogeneous ensemble learning. Boosting, Bagging and Stacking are commonly used ensemble methods [2, 4].

C. Boosting -

Proposed algorithm uses Boosting. In Boosting selection of instances for training is done by their weights. Main idea is to take a weak learner and repeatedly run it, but focus on misclassified examples. Weight of misclassified instances is increased, so it will gain more attention in next iteration. Final classification is based on weighted votes of weak classifiers [5]. Boosting can significantly reduce the bias in addition to reducing the variance, and therefore, on weak learners such as decision stumps, Boosting is usually more effective.

D. Incremental Learning -

It extracts new knowledge from new training data and adds it to previous knowledge [1]. It is also called as relearning. It overcomes limitations on training data we can add more training data whenever it become available, new data can be easily added without needing previous data [9].

E. Objectives -

1. The learning algorithm must be able to learn additional information from new training data.
2. The learning algorithm should not require the previously used training data, which is used to train the existing classifier.
3. The learning algorithm should not forget the previously acquired knowledge.

II. PROPOSED ALGORITHM

Proposed algorithm is based on ensemble learning. Homogeneous ensemble of classifiers is created by using decision stump. Each classifier trained on a subset of the different distributions of the available training dataset, and then combined using weighted majority voting. It creates additional ensemble when new data arrives. In each iteration it creates a new classifier and adds it to the list of classifiers. Once classifier is build it updates the weight of misclassified training instances, so that it will gain more attention in next iteration. It considers entire ensemble build so far to classify instance. In training mode we also calculate weights of the classifiers which are useful for weighted majority voting in classification step. We are using decision stump (weak learner) as a classification algorithm. As we do not discard any previously build classifiers, the previously acquired knowledge is regained and that makes this algorithm to learn incrementally.

A weak learner is a learning algorithm capable of producing classifiers with probability of error strictly less than that of random guessing 0.5. A decision stump is a very simple decision tree. It's a tree with only one split, so it's a stump. A decision stump makes a prediction based on the value of just a single input feature [3]. Sometimes they are also called 1-rules.

The flowchart of proposed algorithm is shown in Figure 1. Following are the inputs to proposed incremental learning algorithm.

1. Training dataset – L
2. A weak learner – W
3. No of iterations to be performed – T (No of classifiers to be build)

Figure 2. Shows the reweight step and Figure 3. Shows the classification step.

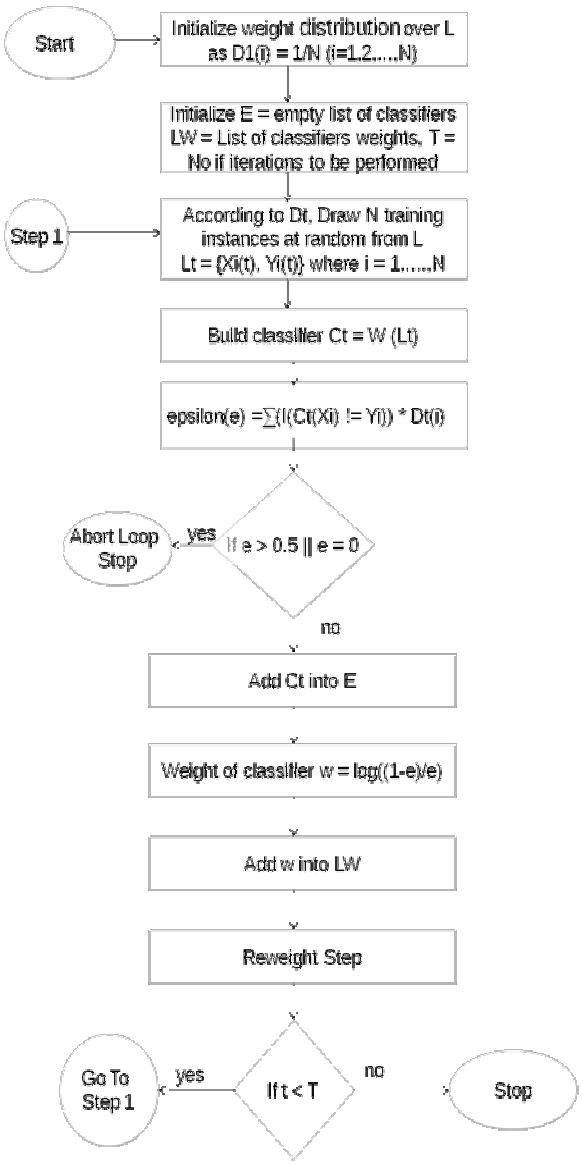

Figure 1. Proposed algorithm

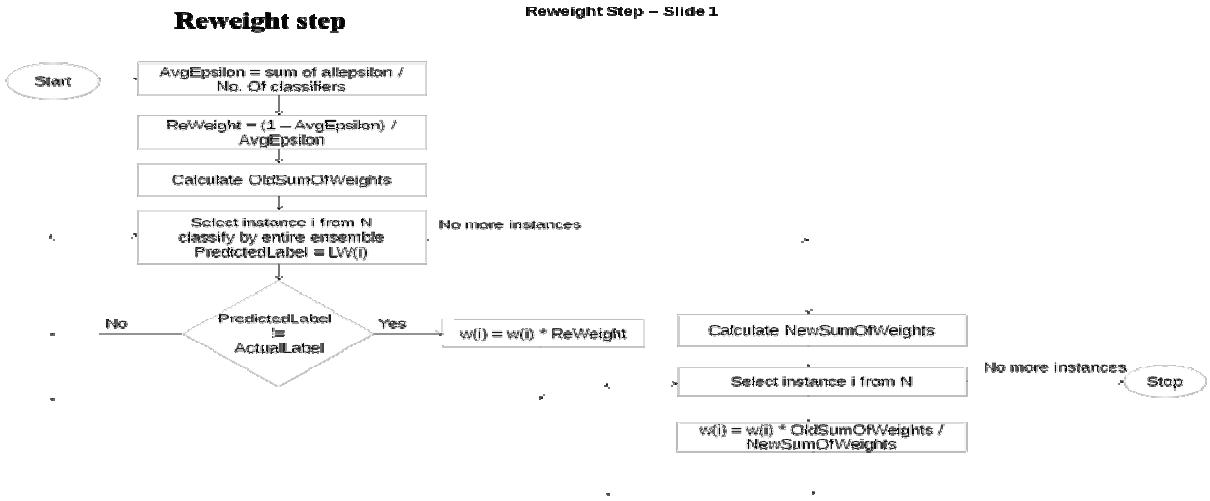

Figure 2. Reweight step

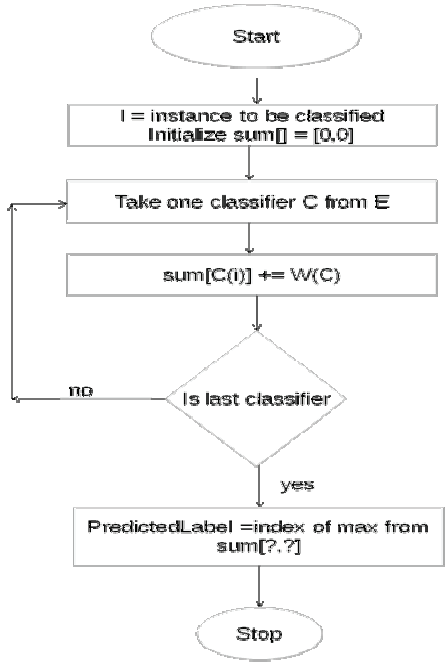

Figure 3. Classification step

III. EXPERIMENT AND RESULT

There are different public datasets specially designed and used for spam filtering algorithms like Spambasse, lingspam, spamassassin, trec05p-1. For evaluation purpose a dataset is divided into a training set and testing set. Each dataset consist of number of emails and each email contains body and header. Preprocessing is applied on email to break it into tokens. Stop words are removed. e.g. eliminate words that are often occurred. Stemming used to find out the root of a word. Stemming converts words to their stems.

Proposed algorithm is tested with various datasets, result of proposed algorithm is compared to other available incremental and non incremental algorithms. Proposed algorithm performs better .when data is incrementally added it shows its incremental ability to achieve high accuracy

According to the Table 1 and Table 2 it is clear that the accuracy of the proposed incremental algorithm is comparatively better than Learn++ and Adaboost algorithm. Other results of proposed algorithm like spam Precision and spam recall is similar and even better than non-incremental algorithm.

Following formulas are used to calculate performance measure

1. Accuracy = No. of emails correctly categorized / total No. of emails
2. Spam Precision = No. of spam correctly classified / total No. of emails classified as spam
3. Spam Recall = No. of spam correctly classified / total No. of emails

Table -1 Experiment Result For LingSpam dataset

Measure	The proposed Algorithm	Learn++ Algorithm	Adaboost Algorithm
Accuracy	94.13	78.64	93.68
Precision	0.91	0.55	0.84
Recall	0.83	1	0.95

Table -2 Experiment Result For SpamAssassin dataset

Measure	The proposed Algorithm	Learn++ Algorithm	Adaboost Algorithm
Accuracy	99.97	99.80	99.98
Precision	0.99	0.99	0.99
Recall	1	1	1

IV.CONCLUSION

The proposed incremental learning algorithm is able to classify a given data into corresponding categories as spam or ham. When new training data is made available, by using Incremental Spam Filter we get really promising results.The experimental results show that our approach has a high Accuracy and which is verified.

REFERENCES

[1] Elham Ghanbari & Hamid Beigy, "An Incremental Spam Detection Algorithm." *Artificial Intelligence and Signal Processing (AISP), 2011 International Symposium on Tehran* .

[2] Devi Parikh and Robi Polikar "An Ensemble-Based Incremental Learning Approach to Data Fusion", *Member, IEEE,IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B: CYBERNETICS, VOL. 37, NO. 2, APRIL 2007*

[3] Wayne Iba Ai , Pat La ngley, "Induction of One-Level Decision Trees (1992)" ,*Ninth International Conference on Machine Learning,1992.*

[4] Chun-Xia Zhang *, Jiang-She Zhang,C.-X. Zhang, J.-S. Zhang "RotBoost: A technique for combining Rotation Forest and AdaBoost", *Pattern Recognition Letters 29 (2008)*

[5] S.B. Kotsiantis, D. Kanellopoulos and P.E. Pintelas "Local Boosting of Decision Stumps for Regression and Classification Problems", *JOURNAL OF COMPUTERS, VOL. 1, NO. 4, JULY 2006*

[6] Phimphaka Taninpong Sudsanguan Ngamsuriyaroj , "Incremental Naïve Bayesian Spam Mail Filtering and Variant Incremental Training", *2009 Eigth IEEE/ACIS International Conference on Computer and Information Science.*

[7] Juan J. Rodrıguez *, Jesus Maudes, J.J. Rodrıguez, J. Maudes , "Boosting recombined weak classifiers", *Pattern Recognition Letters 29 (2008)*

[8] Thiago S. Guzella *, Walmir M. Caminhas , "A review of machine learning approaches to Spam filtering", *Department of electrical engineering, federal university of minas gerais 31270-910 Brazil(2009)*

[9] Robi Polikar, Lalita Udapa, satish Udapa Learn++: An Incremental Learning Algorithm for Supervised Neural Networks, *IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 31, NO. 4, NOVEMBER 2001*