ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

An Effective Algorithm & Comparison of Various Techniques using J48 Classifier

Shiva Sharma [1], U. Datta [2]

P.G. Student, Department of Computer Engineering, MPCT Engineering College, Gwalior, M.P, India[1]

Associate Professor, Department of Computer Engineering, MPCT Engineering College, Gwalior, M.P, India[2]

ABSTRACT: Nowadays a primary trouble in spam filtering in addition to textual content classification in natural language processing is the huge size of vector area due to the several characteristic terms that is commonly the purpose of widespread calculation and slow classification. Support Vector Machine (SVM) takes a set of input data and output the prediction that data lays in one of the two classes i.e. It classify the data into possible classes. SVM has the greater ability to generalize the problem, which is the goal in statistical learning. The statistical learning theory provides an outline for studying the problem of gaining knowledge, making predictions, making decisions from a set of data. In the existing work, Support Vector machine (SVM) used for training and testing datasets. It has many drawbacks which degrades the performance of process. Although SVMs have good generalization performance, they can be abnormally slow in test phase. Another limitation is speed and size, both in training and testing. the feature vector of every email will be extracted by the feature selection module. Because most of the features present redundancy and inconsistency, we adopt a feature selection method that is based on the information gain (IG). Specifically, we compute the IG for every feature vector, no matter whether it corresponds to a spam or a regular email. These feature vectors are then ordered based on their IG values, in a decreasing order.

KEYWORDS: Information gain, Support Vector machine, Spam, e-mails.

I. INTRODUCTION

Mobile SVM, a standout amongst other machine learning algorithms, which was proposed in 1990 and generally utilized for design recognition. Likewise image recognition, speech recognition, text characterization, face detection and faulty card detection and so on like numerous worldview has connected for classification issues. SVM machine learning is a mix-up [1]. In the algorithm that, given an arrangement of preparing cases, each identified with one of the few classifications as, A model that predicts that the new SVM training algorithm builds a scope of illustration. SVM learning for the general issue, which is going for more noteworthy measurable limit. In measurable learning theory the issue of managed learning is figured as takes after. We are given an arrangement of training data {(x1, y1)... (xn, yn)} In Rn x R examined by obscure probability distribution P(x, y), and a misfortune work V(y, f(x)) that measures the bugs, for a given x, f(x) is " predicted " rather than the actual value y. The issue comprises in finding a capacity f that limits the desire of the error on new data i.e. finding a function f that minimizes the expected error: $\int V(y, f(x)) P(x, y)$ dx dy [2] SVM has attracted a great deal of attention in the last decade and actively applied to different space applications. SVMs are regularly utilized for learning characterization, regression or ranking function. SVM depends on factual learning hypothesis and structural risk minimization key and have the point of deciding the area of choice limits otherwise called a hyper plane that deliver the ideal partition of classes. Maximizing the margin and in this manner making the biggest conceivable separation between the separating hyper plane and the instances on either side of it has been demonstrated to lessen an upper bound on the expected generalization error.

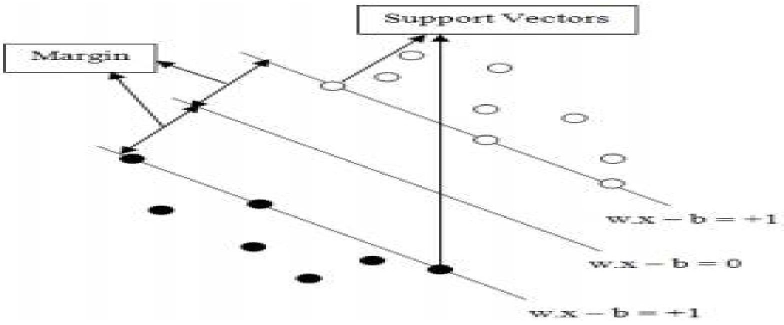


Figure: SVM Model

Related work

Traditional techniques are isolated into the accompanying classifications:

Methods based on analysis of messages: The got e-mail is broke down for particular indications of spam on the base of:

- Formal signs;
- Content utilizing signature in updated database;
- Content applying measurement strategies in view of Bayes theorem;
- Content by methods for utilize SURBL (Spam URL Realtime Block Lists), when run scan for found references in email and their check under base of SURBL. This technique is powerful if rather than notice, the reference of site with notice is situated in e-mail.. **Proposed algorithm**

Incremental learning follows machine learning paradigm where learning process takes place whenever new set of examples appears. The discriminant function is updated to adjust the new definition of data by incrementally learning over a stream of data. Incremental training of SVM involves quickly re-training a SVM after adding a small number of additional training vectors to the training set of an existing SVM. The ability to include additional training data when it becomes available and to relearn from them is a significant feature of incremental learning. Conventional batch learning follows a hypothesis that the sharing of data is identical in training and testing set. This often degrades the filter performance. Spam e-mails are generally characterized as:

(1) Certain types of spam mails appear for a short duration of time. So, spam features change dynamically.
(2) Certain types of spam mails appear at regular interval. This leads to the problem of recurring contexts i.e. concepts may re-appear in future, a special sub type of concept drift.
(3) Other types of spam mails appear continuously.

II. PSEUDO CODE

```
Step:1   Input datasets from the database
Step:2   Perform Pre-processing over datasets (Training and Testing)
Step:3   Execute tokenization, stop word removal and stemming
Step:4   Perform feature selection over the subset of the overall data
Step:5   Testing Phase
             a.   Input testing dataset
             b.   Testing instances contains set of unlabelled incoming spam and legitimate mails
             c.   Classify testing instances
```

DOI: 10.15680/IJIRCCE.2018.0605115

Step:6 Apply J48 Algorithm

Step:7 Create root node and label with splitting attribute

Step:8 D = Database created by applying splitting predicate to D

Step:9 If stopping point reached for this path

Step:10 Stop

III. SIMULATION RESULTS

In the implementation of the proposed work, we used WEKA which show HAM and SPAM. Enron datasets are used for the detailed analysis of the mails. WEKA (Waikato Environment for Knowledge analysis) is a fashionable suite of machine learning software program written in Java, evolved at the University of Waikato, New Zealand. WEKA is unfastened software program available beneath the GNU General Public License. In the first method, traditional batch schooling is completed on J48. All training examples are presented at the same time and resultant support vectors are used to discriminate e-mails from testing sets. J48 is trained incrementally in the second and third approaches.

Table I: Dataset Size

Datasets	Total Mails		Training Set Size		Testing Set Size	
	Spam	Ham	Spam	Ham	Spam	Ham
Enron 1	1500	3672	500	1224	100	250
Enron 2	1496	4361	400	1100	110	330
Enron 3	1500	4012	500	1300	100	270
Enron 4	4500	1500	1000	500	350	100
Enron 5	3675	1500	1225	500	245	100
Enron 6	4500	1500	1000	500	350	100

Spam Precision:

$$SP = \frac{n_{S\to S}}{n_{S\to S} + n_{L\to S}}$$

Spam Recall:

$$SR = \frac{n_{S\to S}}{n_{S\to S} + n_{S\to L}}$$

Legitimate Precision:

$$LP = \frac{n_{L\to L}}{n_{L\to L} + n_{S\to L}}$$

Legitimate Recall:

$$LR = \frac{n_{L\to L}}{n_{L\to L} + n_{L\to S}}$$

Where $n_{S\to S}$ means correctly spam total spam emails $n_{L\to L}$ means correctly categorized total legitimate emails, $n_{L\to S}$ means legitimate emails categorized as spam, $n_{S\to L}$ means spam emails categorized as legitimate.
Table II and III shows the performance measures for different ENRON datasets of Base and Propose. There are 5 parameters taken into consideration such as Spam Precision, Spam Recall, Legitimate Precision, Legitimate Precision and MCC.

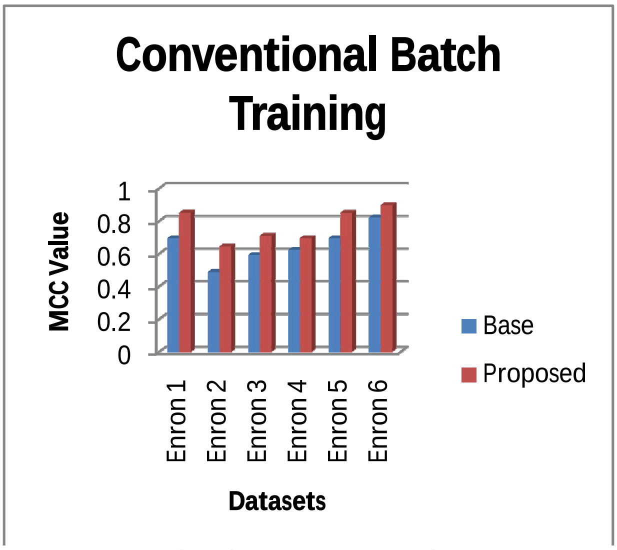


Fig. 5.1 Conventional Batch Training Vs MCC Value

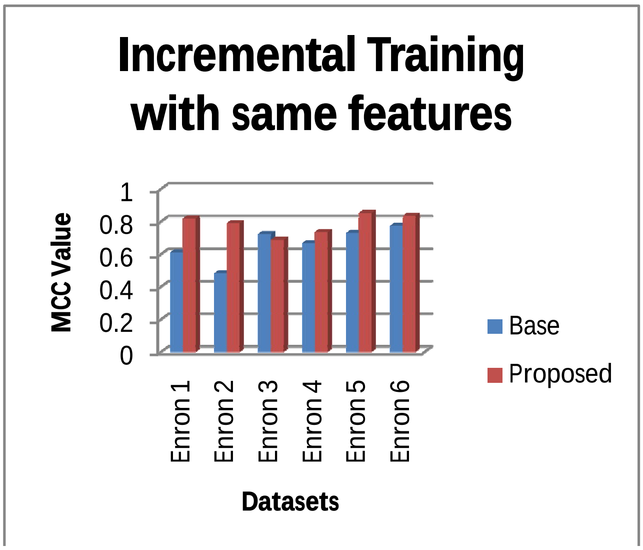
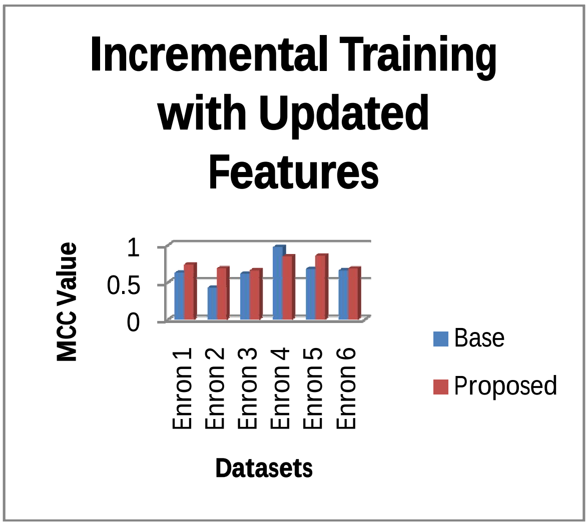


Fig. 5.2 Incremental Training with same features Vs MCC Value



5.3 Incremental Training with Updated features Vs MCC Value

IV. Conclusion and Future Work

The most common scam mails is the fraud job offer mails, most of them are using the logos of multinational companies and higher official names and signatures. The only way to identify the fraud mails and legitimate mails is that the email ids of multinational companies' newer use Gmail, Hotmail or Yahoo, they will have their official mail account. The performance testing on the designed email spam filter is to calculate the accuracy, reliability and other factors. SVM based arrangement is attractive, in light of the fact that its efficiency does not specifically rely upon the measurement of ordered elements. Though SVM is the most powerful and precise grouping method, there are a few issues. Originally, the SVM was created for binary arrangement, and it isn't easy to expand it for multi-class order issue. Spam filtering in Internet email can work at two levels, an individual user level or a venture level. An individual user is ordinarily a person working at home and sending and accepting email by means of an ISP. Such a user who wishes to identify and filter spam email introduces a spam separating framework on her individual PC. Often, image spam contains nonsensical, PC created content which basically irritates the reader. However, new technology in some programs tries to read the images by attempting to find text in these images. They are not extremely exact, and here and there sift through innocent images of items like a container that has words on it.

Social spam is an e-crime on social networking sites with contents such as comments, post, chat, etc. There are many spamming activities going through social media such as malicious links posting, insulting posts, hate speech, fake friends, deceitful reviews, etc. Motivation of these spamming activities can be either private or commercial. Previously, emails were the major object of spammers however it is slowly reduced with the advancement of spam filters that can filter almost 95% of spam content mails. On the other hand, growth of social networking sites and its weak security measures attracted many spammers and made it a vigorous field of concern for research community.

REFERENCES

[1] Ashish Pradhan ‘"SVM-A Survey" International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 8, August 2012).

[2] Ms. Snehal S. Joshi, Mr. Navnath D. Kale “Survey: SVM and Its Deviations in Classification Techniques” Volume 4, Issue 12, December 2014 ISSN: 2277 128X International Journal of Advanced Research in Computer Science and Software Engineering.

[3] J. S. Kong, B. A. Rezaei, N. Sarshar, V. P. Roychowdhury and P. O. Boykin, “Collaborative Spam Filtering Using E-Mail Networks,” IEEE Computer Society on Computer, Vol. 39, No. 8, 2006, pp. 67-73.

[4] A. Gray and M. Haahr, “Personalised, Collaborative Spam Filtering,” Proceedings of the First Conference on Email and Anti-Spam (CEAS), Mountain View, 30-31 July 2004.

[5] R. M. Alguliyev and S. H. Nazirova, “Multilayer and Multiagent Automated Email Filtration System,” Telecommunications and Radioengeneering, Vol. 67, No. 12, pp. 1089-1095

[6] Ms. Ruchida S. Sonar, Dr. P.R. Deshmukh “SVMs for Human Face Detection: A Review” International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169 Volume: 2 Issue: 11.

[7] Omar Saad, Ashraf Darwish and Ramadan Faraj “A survey of machine learning techniques for Spam filtering” IJCSNS International Journal of Computer Science and Network Security, VOL.12 No.2, February 2012.

[8] Saadat Nazirova “Survey on Spam Filtering Techniques” Communications and Network, 2011, 3, 153-160.

[9] M. Kepa, J. Szymanski, "Two stage SVM and kNN text documents classifier," In: Pattern Recognition and Machine Intelligence, Kryszkiewicz M. (Ed.), Lecture Notes in Computer Science, Vol. 9124, pp. 279-289, 2015.

[10] R. C. Barik and B. Naik, "A Novel Extraction and Classification Technique for Machine Learning using Time Series and Statistical Approach," Computational Intelligence in Data Mining, vol. 3, pp. 217-228, 2015.

[11] R. Bruni and G. Bianchi, "Effective Classification Using a Small Training Set Based on Discretization and Statistical Analysis," IEEE Trans. Knowl. Data Eng., vol. 27, no. 9, pp. 2349-2361, 2015.

[12] A. Chaudhuri, "Modified fuzzy SVM for credit approval classification," IOS Press and Authors, vol. 27, no. 2, pp. 189-211, 2014.

[13] E. Baralis, L. Cagliero, and P. Garza, "EnBay: A novel pattern-based Bayesian classifier," Tkde, vol. 25, no. 12, pp. 2780- 2795, 2013.

[14] X. Fang, "Inference-Based Naive Bayes: Turning Naive Bayes Cost-Sensitive," vol. 25, no. 10, pp. 2302-2314, 2013.

[15] C. H. Wan, L. H. Lee, R. Rajkumar, and D. Isa, "A hybrid text classification approach with low dependency on parameter by integrating K-nearest neighbor and SVM," Expert Syst. Appl., vol. 39, no. 15, pp. 11880-11888, 2012.

[16] L. H. Lee, R. Rajkumar, and D. Isa, "Automatic folder allocation system using Bayesian-SVMs hybrid classification approach," Appl. Intell., vol. 36, no. 2, pp. 295-307, 2012.