International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 8, Issue 6, Nov - Dec 2017, pp. 108–120, Article ID: IJARET_08_06_011
Available online at http://www.iaeme.com/ijaret/issues.asp?JType=IJARET&VType=8&IType=6
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
© IAEME Publication

SUPPORT VECTOR MACHINE FOR PERSONALIZED E-MAIL SPAM FILTERING

Gopi Sanghani
Computer Engineering Department, Nirma University,
Ahmedabad, India

Dr. Ketan Kotecha
Parul University, Waghodia, Vadodara, India

ABSTRACT

E-mail is one of the most frequently used personal and official communication tool over the Internet. The continually increasing ratio of spam e-mails over legitimate e-mails and adversarial nature of spam e-mails lead to the requirement of employing spam filter that can be updated dynamically. Moreover, the discrimination criteria of spam and legitimate e-mails vary for different users worldwide. This leads to the personalization of e-mail spam filter which automatically adapts individual user's characteristics. We propose an incremental learning model for personalized e-mail spam filtering. We apply support vector machine - a supervised machine learning algorithm & a discriminative classifier for the designing the classification model. We apply incremental learning using support vector machine for the development of a dynamically updated filter. Our model is evaluated on two different datasets that consist of a set of e-mails structured according to the order of arrival. Experimental results confirm the superior performance of incremental learning over the batch learning model. The inclusion of incremental learning when the distribution of data is different in training and testing sets helps improving classification accuracy and decreases the false positive rate substantially.

Key words: Support Vector Machines, Incremental Training, Personalized Spam Filter, Distribution Shift

Cite this Article: Gopi Sanghani and Dr. Ketan Kotecha, Support Vector Machine For Personalized E-Mail Spam Filtering**.** *International Journal of Advanced Research in Engineering and Technology,* 8(6), 2017, pp 108–120.
http://www.iaeme.com/ijaret/issues.asp?JType=IJARET&VType=8&IType=6

1. INTRODUCTION

Classification of text is the most essential requirement in today's era due to the increasing volume of electronic text over the internet. The huge textual data available from different sources over the net can be used to generate productive outcomes only when appropriate mining or analysis tools are applied to them. Text classification is the process of distributing text documents into two or more predefined classes depending upon the similarity measures of

documents. In automatic content-based text classification, the classifiers learn the class boundary or actual distribution of each class from the training data and classify the unseen samples from the test data. The content-based classification requires the extraction of representative features for the transformation of unstructured text into a more explicit structured format. The performance of such classifier relies upon how efficiently the features are selected to represent the textual data. Moreover, the classifier selection is influenced by the nature of the text data to be classified. It remains static when either all of the sample text is available or the distribution of data does not change over a time. In the case when text data collection is characterized by the inclusion of new or updated data over a time period, the dynamic classification model is required. Classification of textual data in the non-stationary environment remains to be a highly challenging task. It requires addressing the issues like the modification of representative features, dynamic change in training & test data distribution, consistently maintaining the performance of classification model etc…

Machine learning algorithms are extensively used for automatic text classification. Support vector machine (SVM) [1, 4] is a supervised machine learning algorithm works on the structural risk minimization principle from statistical learning theory. SVM is a discriminative classifier that learns the decision boundary between classes and classifies unknown examples using the learned hypothesis. Joachims [2] analyzed how the statistical properties of text classification task and generalization performance of SVM are connected in a quantitative way. The author presented the details of how and why SVMs can achieve good classification performance despite the high dimensional feature space in text classification. SVM's robustness is improved during the learning procedure which tightens the decision boundaries for classification [3]. SVM is a maximal margin classifier; it learns a decision boundary during training which maximizes the distance between samples of the two classes. The design of efficient incremental SVM learning and a detailed analysis of convergence and algorithmic complexity of incremental SVM learning is carried out by Laskov et al. [5]. SVM's ability to incrementally learn with a new set of examples and support vectors is addressed by Syed et al. [6].

In this research, we present the design and development of incremental classifier using support vector machine for personalized e-mail classification. We analyzed the performance of the proposed model for the classification of personalized e-mails organized in chronological order. E-mail is the most useful and reliable tool for a business and personal communication worldwide. The e-mail services face the inevitable downside of a number of unsolicited bulk messages known as spam with the high circulating volume and offensive content. A huge number of spam e-mails delivered every day regardless of the commercial or personal level of interest that delay internet traffic and degrade many on-line services. Personalized e-mail spam filter is the most prevalent application of automated binary text classification problems. The binary text classification problem is defined as: Given a training set of n labeled sample e-mails $T = \{t_1, t_2, \ldots, t_n\}$ and $C = \{c_l, c_s\}$ denotes e-mail categories: legitimate and spam. The task is to learn classification model, which classifies previously unseen e-mails into one of the two categories based on their content. Personalized e-mail spam filtering task requires addressing two major challenges: ever increasing ratio of unwanted and useless spam e-mails and continuously differing context and content of spam e-mails. our research work addresses these issues by developing an incremental classifier using SVM and dynamically regenerating the set of representative features. The rest of the paper is organized as follows: Section 2 discusses the review of literature which provides the strong basis for our research. Section 3 describes the details of design and development of incremental classifier using SVM. In Section 4, we present experimental results of two real-world datasets that prove the efficiency of our system. We conclude in Section 5 with an insight into the prospective significance of our results and scope for future work.

2. LITERATURE REVIEW

E-mail spam filters use different approaches to detect spam messages and categorizing e-mails into separate folders. Text classification task considers an approach based on the analysis of message content of an e-mail. For a personalized e-mail spam filtering, the filter is employed by a single user as a client side filter and messages identified as spam are usually sent to spam folder. Filtron [7], a personalized anti-spam filter based on machine learning text categorization paradigm, had been evaluated in real life scenario that confirmed the prominent role of machine learning techniques for anti-spam filtering. Cheng & Li [8] proposed combined supervised and semi-supervised classifier using SVM for personalized spam filtering. Chang, Yih, & McCann [9] designed a light-weight user model that is highly scalable and can be easily combined with a traditional global spam filter. A personalized spam filter is presented by Junejo & Karim [10] using an automatic approach which built a statistical model of spam and non-spam words from the labeled training dataset. The filter is updated in two passes over unlabeled samples taken from individual user's inbox.

Ghanbari & Beigy [11] proposed the algorithm called incremental RotBoost, an incremental learning algorithm based on ensemble learning. Hsiao and Chang [12] developed an incremental cluster-based classification method, called as ICBC. It runs in two phases. The first phase performs clustering of e-mails in each given class into several groups, and an equal number of features are extracted from each group. The second phase includes an incremental learning mechanism for ICBC so that it can adapt itself to accommodate the changes of the environment in a fast and low-cost manner. Georgala, Kosmopoulos, and Paliouras [13] proposed active learning approach using incremental clustering for spam filtering. Taninpong and Ngamsuriyaroj [14] proposed an incremental spam mail filtering using Naïve Bayesian classification in which the sliding window concept is applied to keep the training set to a limited size and the training set is updated when new emails are received.

3. DESIGN & DEVELOPMENT OF INCREMENTAL FILTER

Automatic text classification model requires a set of training samples from which the classifier learns the statistical distribution of data. Traditional server based e-mail spam filters use generic mail corpus for the training and then, commonly applied to a user's inbox to discriminate spam and legitimate e-mails. Worldwide internet users have highly dissimilar perceptions about the definition of e-mail spam, where only global filters may not offer an acceptable performance, as the statistical property of feature space is derived commonly [15]. The end user remains highly dependent on the discrimination of e-mails characterized by the general training corpus. The essential advantage is users are relieved from the burden of processing thousands of unsolicited e-mails. But only global filters cannot optimally reflect individual user's characteristics while discriminating e-mails. As an extensive model, personalization of e-mail spam filtering is required which facilitate robustness and should be adaptive to individual user's preferences. Moreover, the content of spam e-mail changes as spammers continuously change the manner to present the content of spam e-mails. So, there is a need to update the filter dynamically to handle the changing distribution of representative features. The algorithm for the proposed incremental support vector learning model for the personalized e-mail spam filter is shown in table 1 and a detailed explanation is given in the subsequent subsections.

3.1. Classification using SVM

SVM is a supervised machine learning algorithm essentially used for binary classification problems. In binary classification problem, a data set X contains *n* labeled example vectors $\{(x_1, y_1) \ldots (x_n, y_n)\}$. Here x_i represents the input vector with corresponding binary labels

denotes as $y_i \in \{-1, 1\}$. Let $\varphi(x_i)$ be the corresponding vectors in feature space, where $\varphi(x_i)$ is the implicit kernel mapping such that $k(x_i, x_j) = \varphi(x_i) \cdot \varphi(x_j)$ be the kernel function, implying a dot product in the feature space. The optimization problem for a soft-margin SVM is,

$$\min_{w,b} \left\{ \frac{1}{2} ||w||^2 + C \sum_i \xi i \right\} \quad (1)$$

Subject to the constraints $y_i (w.x + b) = 1 - \xi_i$ and $\xi_i \geqslant 0$ where w is the normal vector of the separating hyperplane in feature space and $C > 0$ is a regularization parameter controlling the penalty for misclassification. Equation (1) is referred to as the primal equation. From that, the Lagrangian form of the dual problem is:

$$w(\alpha) = \max_{\alpha} \left\{ \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \, k(x_i, x_j) \right\} \quad (2)$$

Subject to $0 \leqslant \alpha_i \leqslant C$. This is a quadratic optimization problem that can be solved efficiently using algorithms such as Sequential Minimal Optimization [16]. Many α_i go to zero during optimization and the remaining x_i corresponding to those $\alpha_i > 0$ are called support vectors. If l is the number of support vectors and $\alpha_i > 0$ for all i, with this formulation, the normal vector of the separating plane w is calculated as:

$$w = \sum_{i=1}^{l} \alpha_i \, y_i x_i \quad (3)$$

The classification f(x) for a new sample vector x can be determined by computing the kernel function of x with every support vector:

$$f(x) = \text{sign}\left(\sum_{i=1}^{l} \alpha_i \, y_i . k(x, x_i) + b \right) \quad (4)$$

Here the bias term b is the offset of the hyperplane along its normal vector, determined during SVM training. SVM algorithm maps input vectors into a feature space of higher dimension and constructs an optimized hyperplane for generalization. The training samples lying near to the hyperplane are called support vectors.

3.2. Incremental model for Personalized E-mail Spam Filter

Personalized e-mail spam filter serves as an extensive model whenever the users tend to differ in their interest and preferences for discrimination of e-mails. Generally, user's preferences are influenced by their personal interests, professional profile, hobbies, etc…Many content-based spam filters apply machine learning techniques, of which support vector machine has shown consistently superior performance. SVM was initially applied for spam categorization by Drucker, Wu, and Vapnik [17]. Since then various extensions and approaches based on online and active learning have been presented by many researchers because of SVM's good generalization ability and higher classification accuracy.

Table 1 Personalized e-mail Spam Filter using Incremental Support Vector Learning

	Input: Training Set $Trem_0 = \{Em_s\} \cup \{Em_l\}$ $\varrho \leftarrow$ threshold value for Accuracy $Trem_0 \leftarrow$n training e-mails with labels $Em_s \leftarrow$ Set of Spam e-mails $Em_l \leftarrow$ Set of Legitimate e-mails
1	***Pre-processing of training & Testing sets $Trem_0$&Tsem*** Tokenization Stop word removal Stemming
2	***Feature Selection*** Generate a subset of Representative features FS_i using Information Gain (IG) feature selection method Represent e-mail messages using vector space model (VSM) i = 0
3	***Build a classifier: SVM Conventional Training using SMO Algorithm (Pass I)*** Output: Support Vector Set $\alpha_i = \{\alpha_k \mid k = 1 \text{ to } l\}$

4	***Testing Phase (Pass II)***
	Input: Testing Sets Tsem = $\{Ts^1, Ts^2, \ldots, Ts^m\}$
	// Testing instances contains set of unlabelled incoming spam and legitimate e-mails
	Repeat
	Classify testing instances $Ts^1, Ts^2, \ldots$ from Tsem
	until either accuracy $\leq \varrho$ or FPR increases
	
	$i = i + 1$

5	***Incremental model using SVM with updated feature set (Pass III)*** Input: Resulting Support Vector Set α_{i-1} Re-training set $Rtrem_i = \alpha_{i-1} \cup Ts^k$, where Ts^k is the testing instance for which accuracy / FPR constraint is violated.
5.1	Update the set of Representative Features as follows, $FS_i = FS_{i-1}$. Generate new subset of Features NFS_i from $Rtrem_i$ Update the feature set FS_i as follows: for each **distinct** feature *l* in NFS_i if IG_SCORE($feature_l$) > Avg(IG_SCORE(FS_{i-1})) then $FS_i = FS_i \cup \{feature_l\}$ & $FS_i = FS_i - \{feature_q \mid feature_q$ has lowest IG SCORE$\}$
5.2	***Retrain SVM on the re-training set $Rtrem_i$*** Output Support Vector Set $\alpha_i = \{\alpha_k \mid k = 1 \text{ to } l\}$ Repeat step 4 {Testing Phase} with $k = k + 1$ and step 5

Our classification algorithm is developed as follows: the incremental filtering process is carried out over three passes. The first pass (Pass I in table 1) is performed using conventional batch training of SVM, with n labeled examples, that generates the discriminant function F(x). Pass II comprises a series of testing phases in which small batches of incoming unlabeled e-mails are given to identify true labels. Pass III is carried out by activating incremental training whenever any one of the two performance criteria – accuracy and false positive (FP) rate are violated. Whenever an accuracy of the filter decreases below some threshold or the false positive rate increases, the incremental training will be initiated. Accuracy decreases when miss-classified mail ratio increases, which indicates that the re-training is required to upgrade the filter performance. False positive (FP) rate is also more sensitive because the higher rate of recognizing non-spam messages as spam would definitely degrade the filter performance. Decreasing accuracy or increasing FP rate is the indication of an increase in the error of classification model. A substantial increase in the error of the algorithm because of a change in the class distribution signifies that the current decision model has become less effective over a certain period of time.

An important property SVM possesses is, a set of support vectors represents the feature space and class boundaries in a very concise manner. So, incremental SVM can be trained by preserving support vectors and adding them to the next batch of incoming examples. Initially, conventional batch training is conducted with the substantial number of labeled e-mails representing both spam and legitimate categories. In case of personalized spam e-mails, certain types of spam e-mails appear for a short duration of time while some of the spam e-mails appear regularly. The statistical properties of training dataset and testing datasets differ whenever a class or data distribution changes. Either the target concept or the data distribution change over a time, which may often lead to the necessity of revising the current model. Individual user's preferences for unwanted messages may remain same over a long period of time; the relative frequency of different types of spam may change drastically with time. So, before activating incremental training feature space is updated dynamically to include new features with higher discriminating ability. The feature set is updated using a heuristic function which re-calculates feature's information gain score from re-training set before activating the incremental training. Modifying the feature set would enable the classifier to effectively re-learn the updated distribution of data. The change in the content of e-mails or change in the preferences of user apparently requires updating the filter dynamically. During incremental training, the true label of e-mails is provided in order to correctly derive the modified hypothesis function. The re-training results in modified discriminant function which enables the classifier to handle the distribution shift proficiently.

4. EXPERIMENTS AND RESULTS

We have applied Sequential Minimal Optimization (SMO), a special case of decomposition method for SVM training. The SMO algorithm is designed to avoid the large quadratic optimization problem that is required to solve for the implementation of SVM classification model. At every step, the SMO algorithm analytically solves a QP sub-problem for the two chosen Lagrange multipliers and updates the SVM model accordingly. SMO maintains kernel matrix of size equal to the total number of samples in the dataset, which allows it to handle very large training sets. Incremental SMO learning is achieved by keeping the old α value of support vectors and setting α value to zero for new examples.

4.1. Dataset and Experiments

To validate the accuracy of the proposed incremental model, we use two different datasets Enron [18] and ECUE [19]. Both the benchmark datasets contain the personalized collection

of e-mails. Another essential property the datasets possesses is, the e-mails are as per the order of arrival which perfectly suits the evaluation of our incremental model. The first well-known dataset Enron contains six large personalized folders of spam and legitimate e-mails. This dataset contains pre-processed e-mail messages with the removal of attachments. The dataset belongs to six e-mail directories farmer-d, kaminski-v, kitchen-l, williams-w3, beck-s, and lokay-m, named as Enron1 to Enron6. Text pre-processing tasks such as tokenization, stop word removal and stemming are performed using Rapid Miner [20] prior to applying filtering process. Processed e-mails are represented using Vector Space Model (VSM) in which each e-mail is represented by n dimension vector using binary representation. In three of the folders, a legitimate-spam ratio is approximately 3:1, while in the other three the ratio is inverted to 1:3. The total number of messages in each dataset is between five and six thousand. SVM is trained on a set of e-mails taken from individual user's inbox to capture personalization. In the simulation run, SVM is initially trained with an approximately 33% e-mails including legitimate and spam both. Remaining e-mails are used to create incoming testing instances. As the dataset contains chronologically sorted e-mails, ten different testing instances of equal size are created. The distribution of e-mails for training and testing sets is given in table 2 for Enron dataset and in table 3 for ECUE dataset.

Table 2 Enron Dataset Distribution

Dataset		**Training E-mails**		**Testing E-mails**	
		Legitimate	**Spam**	**Legitimate**	**Spam**
ENRON	ENRON1	1224	500	1000	2448
	ENRON2	1100	400	1096	3261
	ENRON3	1300	500	1000	2712
	ENRON4	500	1000	3500	1000
	ENRON5	500	1225	2450	1000
	ENRON6	500	1000	3500	1000

Table 3 ECUE Test Dataset Distribution

Month	**CDDS1**		**CDDS2**	
	Legitimate	**Spam**	**Legitimate**	**Spam**
Feb'03	--	--	151	142
Mar	93	629	56	391
Apr	228	314	144	405
May	102	216	234	459
Jun	89	925	128	406
Jul	50	917	19	476
Aug	71	1065	30	582
Sep	145	1225	182	1849
Oct	103	1205	123	1746
Nov	85	1830	113	1300
Dec	105	576	99	954
Jan '04	--	--	130	746

The ECUE 1 and ECUE 2 datasets are taken from the ECUE concept drift 1 and 2 datasets, respectively. Each dataset is a collection of e-mails received by the individual user over the period of 10 to 12 months. These dataset contains three types of features: (a) word features, (b) letter or single character features, and (c) structural features, e.g., the proportion of uppercase

or lowercase characters. In this dataset, a separate set of training e-mails is given which contains 500 spam and 500 legitimate e-mails. And testing e-mails contain a total of approximately 10,000 e-mails that are separated month wise as shown in table 3. The organization of Enron and ECUE allow our incremental model to retrain the classification model using a small set of new examples in order to update the filter dynamically whenever validation criteria are violated.

4.2. Performance Measures

The filter is evaluated with well-known performance measures used in classification. We measure accuracy, false positive rate and false negative rate defined as:

$$Accuracy = (|n_{l\to l}| + |n_{s\to s}|)/(N_L + N_S) \quad (5)$$

$$false\ positive\ rate\ (FPR) = |n_{l\to s}|/N_L \quad (6)$$

$$false\ negative\ rate\ (FNR) = |n_{s\to l}|/N_S \quad (7)$$

Where, NS and NL are total spam and legitimate e-mails, nl→l and ns→s are the numbers of legitimate and spam e-mails classified correctly and nl→s and ns→l are the numbers of legitimate and spam e-mails not correctly classified. The other two success measures employed are micro-F1 and macro-F1 defined as:

$$Micro - F1 = (2 \times P \times R)/(P + R) \quad (8)$$

$$Macro - F1 = \sum_{k=1}^{C} F_k/C \ \ where\ F_k = (2 \times P \times R)/(P + R) \quad (9)$$

Where P and R denote precision and recall measures given as:

$$precision\ (P) = (n_{s\to s})/(n_{s\to s} + n_{l\to s}) \quad (10)$$

$$recall\ (R) = (n_{s\to s})/(n_{s\to s} + n_{s\to l}) \quad (11)$$

4.3. Results and discussion

In this research, an experiment is carried out with an aim of analyzing and comparing the performance of batch learning model with incremental learning using SVM in dynamic feature space. Also, we analyze the performance of incremental learning model in the presence of distribution shift. Generally, training data defines distribution and derives discriminant function, which performs well when testing data follows the same distribution. In the case of a distribution shift, the discriminant function has to be updated to maintain and improve the performance. Moreover, the changes in data distribution causes the requirement of modifying the set of representative features to precisely define the class boundary. The selection of representative features is done with Information Gain [21] feature selection method.

Table 4 shows accuracy, precision, and recall, the most common performance measures in binary classification for Enron and ECUE datasets. The precise comparison results of batch training and incremental training models clearly indicates the superior performance of the incremental model in all the cases. Fig. 1 shows the comparison graph for the accuracy achieved for both the models. As the testing data is sorted chronologically, we have created ten testing instances in Enron dataset. In ECUE, the testing instances are already given month wise. We observe that in the case of conventional SVM training, the accuracy decreases from the testing set TS1 to TS10. In conventional training, SVM is trained initially once only. E-mail datasets are sorted as per arrival order so we can say that over a period of time due to the changing nature of e-mails the classification model becomes ineffective. The classification model consistently performs well by improving the accuracy level using incremental re-

training. The accuracy is averaged over all testing instances. Updating the set of representative features before re-training SVM allows the classification model to relearn the modified distribution of data.

Table 4 The classification results for Personalized e-mail spam filtering

Datasets	Inbox	SVM Incremental Training with updated Feature Set			SVM Batch Training		
		Accuracy	Precision	Recall	Accuracy	Precision	Recall
ENRON	ENRON1	**96.28**	**0.97**	**0.95**	94.18	0.95	0.93
	ENRON2	**96.01**	**0.98**	**0.90**	87.90	0.93	0.62
	ENRON3	**96.76**	**0.98**	**0.93**	93.83	0.96	0.89
	ENRON4	**98.19**	**0.98**	**0.98**	96.49	0.98	0.96
	ENRON5	**97.45**	**0.94**	**0.99**	96.14	0.90	0.99
	ENRON6	**96.86**	**0.94**	**0.98**	91.58	0.92	0.92
ECUE	ECUE 1	**97.32**	**0.99**	**0.96**	91.63	0.98	0.92
	ECUE 2	**96.52**	**0.99**	**0.92**	86.18	0.99	0.85

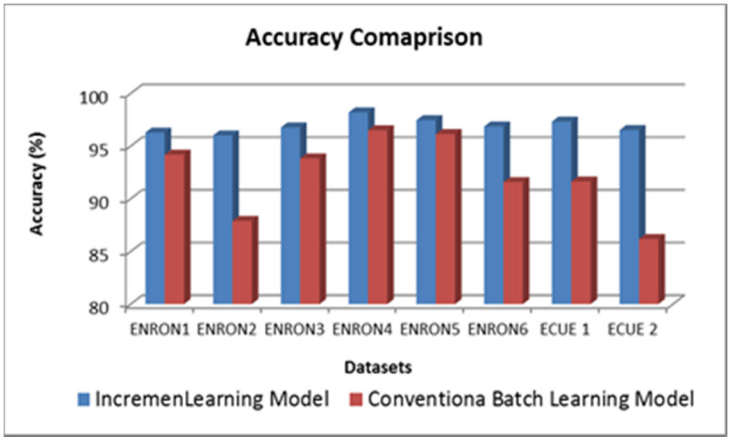

Figure 1 Accuracy Comparison for Batch and Incremental Learning model

Fig. 2 shows FPR comparison for both the learning models. Incorrect classification of legitimate e-mails as spam *i.e.* the occurrence of False Positives (FP) degrades the filter performance. An FP is significantly more harmful than a False Negative (FN) *i.e.* a spam e-mail incorrectly classified as legitimate. Fig. 3 shows the comparison of the false negative rate. Fig. 4 shows ROC curve for the comparison of true positive rate (TPR) vs. false positive rate (FPR). Classification results show that incremental training of SVM allows obtaining and substantially improving the accuracy of the filter. Moreover, the average FPR achieved in the incremental model is decreased by 34% as compared to the average FPR in the batch model. Fig. 5 and 6 show the comparison of Micro F1 measure and Macro F1 measures achieved in both Enron and ECUE datasets respectively. The incremental model successfully handles the change of data distribution and improve the filter performance noticeably.

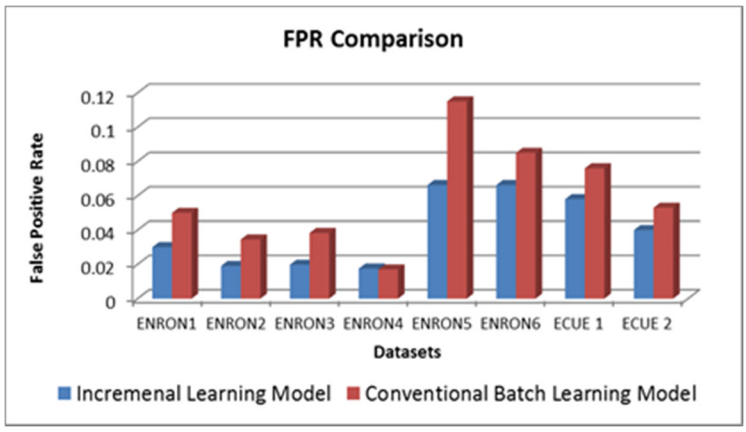

Figure 2 False Positive Rate Comparison for Batch and Incremental Learning model

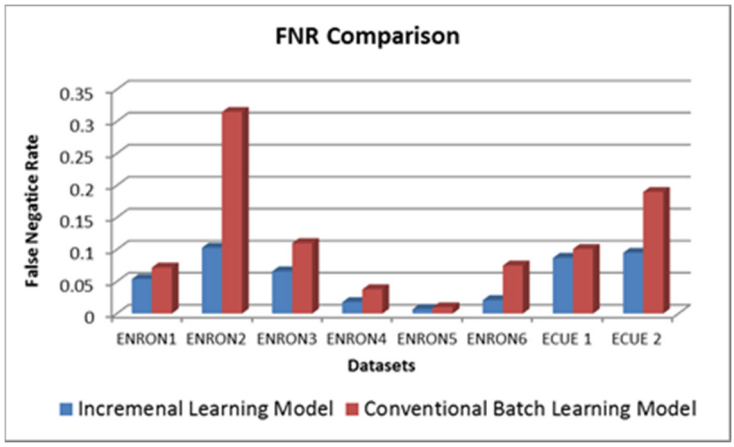

Figure 3 False Negative Rate Comparison for Batch and Incremental Learning model

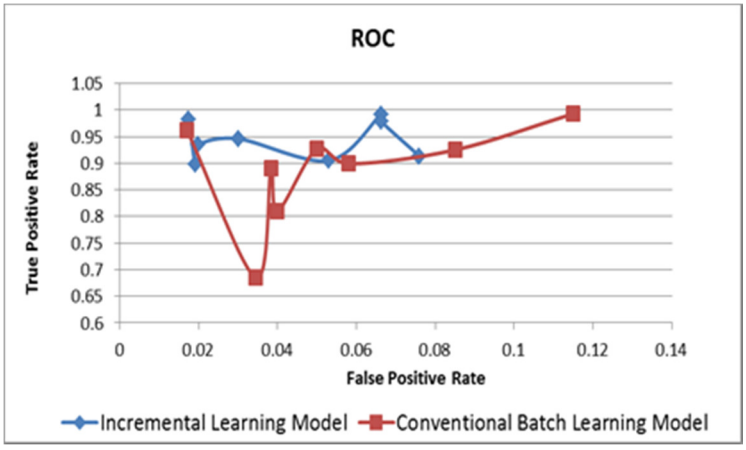

Figure 4 ROC Comparison for Batch and Incremental Learning model

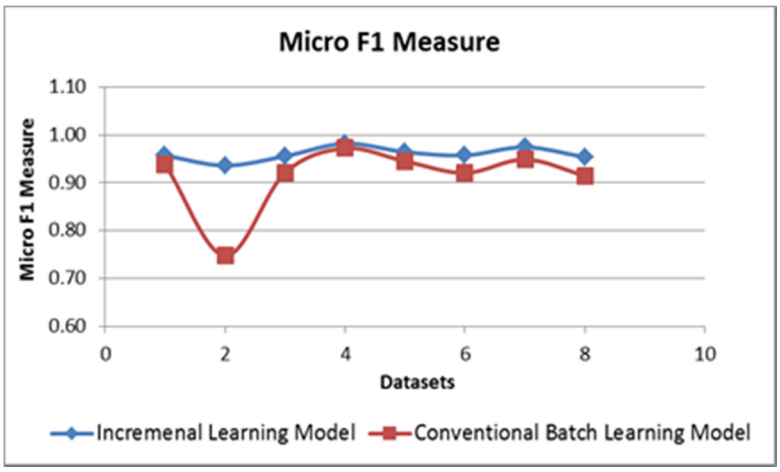

Figure 5 Micro F1 Comparison for Batch and Incremental Learning model

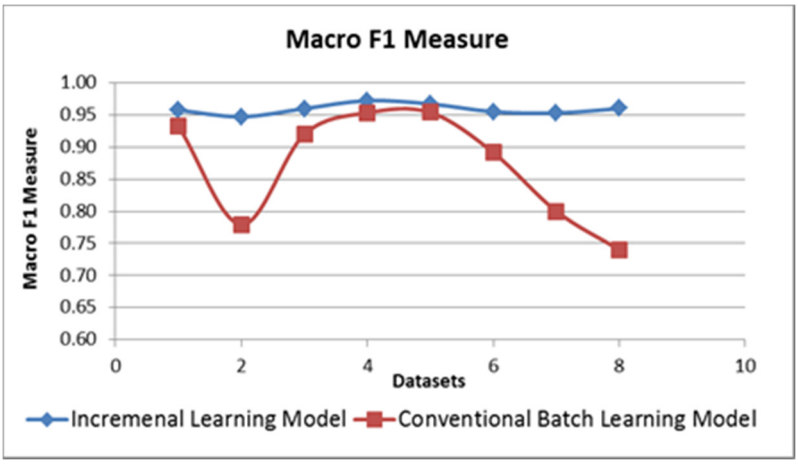

Figure 6 Macro F1 Comparison for Batch and Incremental Learning model

5. CONCLUSIONS

E-mail spam filtering on a personalized level has been one of the most challenging classification tasks in the presence of distribution shift. In this paper, we describe the design, development and evaluation of a personalized e-mail spam filter using incremental training of support vector machine. We apply the technique for the modification of representative features before initiating the incremental training. The experimental outcomes show that the incremental learning effectively helps the classification model to re-learn the modified class boundary information. Hence, this validates the purpose of the application by successfully addressing two major issues of personalized e-mail spam filtering i.e. changing user preferences and distribution shift. SVM incremental learning algorithm outperforms the batch model and the false positive rate is decreased by 34%. The results confirm the applicability of our unique approach that focuses on the incremental learning of SVM with the heuristically updating feature set for the improvement of the classification model. The future work addresses to resolve the issue of dynamic changes in the set of representative features for the prediction of a distribution shift.

ACKNOWLEDGEMENT

We are grateful to the Nirma University for providing resources and other facilities to carry out this research work.

REFERENCES

[1] Cortes, C., & Vapnik, V. Support-vector networks. *Machine Learning*, **20**, pp. 273–297, 1995.

[2] Joachims, T. A statistical learning model of text classification with Support Vector Machines. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, Proc. SIGIR-01, 24th ACM International Conference on Research and Development in Information Retrieval, pp. 128–136, 2001.

[3] Tu., Z. Learning generative models via discriminative approaches. Proceedings CVPR, 2007.

[4] Vapnik, V. Statistical Learning Theory. Wiley, Chichester, GB, 1998.

[5] Laskov, P., Gehl, C., Krüger, S., Müller, K.R. Incremental support vector learning: analysis, implementation and applications. *J Mach Learn* Res **7**, pp. 1909–1936, 2006.

[6] Syed, N., Liu, H., & Sung, K. Incremental learning with support vector machines. Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence, 1999, Stockholm, Sweden.

[7] Michelakis, E., Androutsopoulos, I., Paliouras, G., Sakkis, G., & Stamatopoulos, P. Filtron: a learning-based anti-spam filter, Proceedings of 1st Conf. on Email and Anti-Spam, 2004.

[8] Cheng, V., & Li, C. Personalized spam filtering with semi-supervised classifier ensemble, in WI-06: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, IEEE Computer Society, 2006, pp. 195–201.

[9] Chang, M., Yih, W., & McCann, R. Personalized spam filtering for gray mail. In CEAS-08: Proceedings of 5th Conference on Email and Anti-Spam, 2008.

[10] Junejo, K., & Karim, A. Robust personalizable spam filtering via local and global discrimination modeling. *Knowledge and Information Systems*, **34 (**2), 2013, pp. 299-334.

[11] Elham Ghanbari & Hamid Beigy. Incremental RotBoost algorithm: An application for spam filtering. *Journal Intelligent Data Analysis archive*, **19** (2), 2015, 449-468, IOS Press Amsterdam, The Netherlands.

[12] Hsiao, W.F., & Chang, T.M. An Incremental Cluster-Based Approach to Spam Filtering. *Expert Systems with Applications*, **34** (3), 2007, 1599-1608.

[13] Kleanthi Georgala, Aris Kosmopoulos, & George Paliouras. Spam Filtering: An Active Learning Approach using Incremental Clustering. WIMS Thessaloniki, 2014, Greece Copyright is held by the owner/author(s). Publication rights licensed to ACM.

[14] Phimphaka Taninpong & Sudsanguan Ngamsuriyaroj. Incremental Adaptive Spam Mail Filtering Using Naïve Bayesian Classification. Proc. 10th ACIS International Conference on Software Engineering, Artificial Intelligences, Networking, and Parallel/Distributed Computing, 2009, pp. 243-248.

[15] Guzella T., Caminhas, W. A review of machine learning approaches to spam filtering, *Expert Systems with Applications*, **36 (**7), 2009, pp. 10206–10222.

[16] Platt, J. Fast training of SVMs using Sequential Minimal Optimization. Advances in Kernel Methods Support Vector Machine, MIT Press, Cambridge; 1999, pp.185-208.

[17] Drucker, H., Wu, D., & Vapnik, V. Support vector machines for spam categorization. *IEEE Transactions on Neural networks*, **10** (5), 1999, pp. 1048-1054.

[18] Enron Spam Data sets: http://csmining.org/index.php/enron-spam-datasets.html; 2006.

[19] Delany, S. J., Cunningham, P., & Coyle, L. An assessment of case-based reasoning for spam filtering. *Journal of Artificial Intelligence Review*, 2005, **24** (3-4), pp. 359–378.

[20] RapidMiner: https://rapidminer.com/.

[21] Yang, Y., Pedersen, J. A comparative study on feature selection in text categorization. In D. H. Fisher, editor, Proceedings of ICML-97, Morgan Kaufmann Publishers, San Francisco, US; pp. 412-420