ISSN:0975-9646

A Survey: Classification of E-mail Data Using Semi Supervised Learning

Hiral Dilipbhai Padhiyar[1], Prof. Purvi Rekh[2]

[1]*Department of Computer Engineering, Uka Tarsadia University*
[2]*Department of Computer Engineering, SVNIT*

Abstract: **With the development of Internet and the emergence of a large number of text resources, the automatic text classification has become a research hotspot. Emails is one of the fastest and cheapest communication ways that today it has become the part of communication means of millions of people. It has become a part of everyday life for millions of people, changing the way we work and collaborate. The large percentage of the total traffic over the internet is the email. Email data is also growing rapidly, creating needs for automated analysis. In many security informatics applications it is important to detect deceptive communication in email. As number of training documents increases, accuracy of Text Classification increases. Traditional classifiers (Supervised learning) use only labeled data for training. Labeled instances are often difficult, expensive, or time consuming to obtain. Meanwhile unlabeled data may be relatively easy to collect. Semi-Supervised Learning makes use of both labeled and unlabeled data.**

In the iterative process in the standard EM-based semi-supervised learning, there are two steps: firstly, use the current classifier constructed in the previous iteration to predict the labels of all unlabeled samples; then, reconstruct a new classifier based on the new training samples set. However, there is a problem in the process of reconstructing the training samples. Some unlabeled samples are misclassified by the current classifier because the initial labeled samples are not enough. In this work, an EM based Semi-Supervised Learning algorithm using Naïve Bayesian is proposed in which unlabeled documents are divided into two parts, reliable and misclassified. An Ensemble technique is used to add only reliable unlabeled documents to the training set. Also preprocessing of unlabeled documents is performed before learning process of Naïve Bayesian and SVM classifiers during first step of EM to reduce time of preprocessing, so with this proposed work accuracy of classifier will be increased and execution time will be decreased.

Keywords: Semi supervised learning, classification.

1. INTRODUCTION

Text Categorization (TC) has become one of the major techniques for organizing and managing online information. Several studies proposed the so-called associative classification for databases and few of these studies are proposed to classify text documents into predefined categories based on their contents. The documents to be classified may be texts, images, music, etc. Each kind of document possesses its special classification problems. To automate document classification, a general procedure is as follows: First, a set of pre-classified documents is taken as the training set. The training set is then analyzed in order to derive a classification scheme. Such a classification scheme often needs to be refined with a testing process. The so-derived classification scheme can be used for classification of other on-line documents.

Emails is one of the fastest and cheapest communication ways that today it has become the part of communication means of millions of people. It has become a part of everyday life for millions of people, changing the way we work and collaborate. E-mail is not only used to support conversation but also as a task manager, document delivery system and archive. The downside of this success is the constantly growing volume of e-mail we receive. Against these advantages, some unwanted emails have been created which are called spam [10]. So, Spam is unsolicited and unwanted email from a stranger that is sent in bulk to large mailing lists, usually with some commercial nature sent out in bulk [9].

E-mail users spend and increasing amount of time reading message and deciding whether they are spam or not and categorizing them into folders. E-mail service providers would like to relieve users from this burden by installing server-based spam filters that can classify e-mails as spam automatically [9].

2. LEARNING METHODS

The Machine Learning field evolved from the broad field of *Artificial Intelligence*, which aims to mimic intelligent abilities of humans by machines. Data and Knowledge Mining is learning from data. In this context, data are allowed to speak for themselves and no prior assumptions are made. The step of learning is illustrated in below figure. In step 1, a learning algorithm uses the training data to generate a classification model. This step is also called the **training step** or **training phase**. In step 2, the learned model is tested using the test set to obtain the classification accuracy. This step is called the **testing step** or **testing phase**.

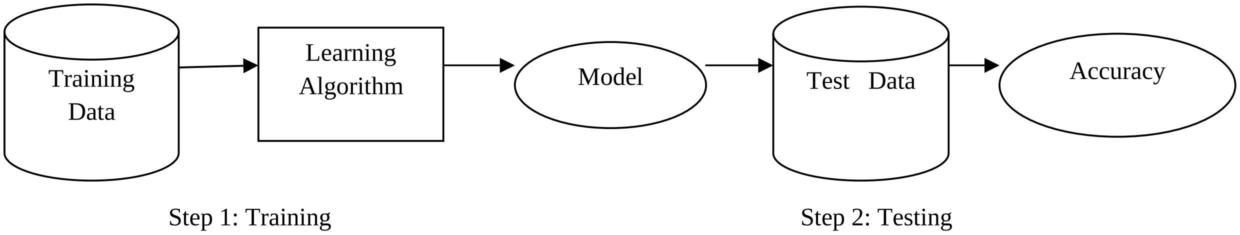


Figure 2.1: The basic learning process: training and testing

If the accuracy of the learned model on the test data is satisfactory, the model can be used in real-world tasks to predict classes of new cases (which do not have classes). If the accuracy is not satisfactory, we need to go back and choose a different learning algorithm and/or do some further processing of the data (this step is called **data pre-processing**, not shown in the figure).

There are mainly 3 types of learning methods.

- Supervised learning
- Unsupervised learning
- Semi-supervised learning

2.1 SUPERVISED LEARNING:

Supervised learning is a Learning based on training data (labeled data). Training Data are sample from the data source with the correct classification/regression solution already assigned. Supervised learning entails learning a mapping between a set of *input* variables *X* and an *output* variable *Y* and applying this mapping to predict the outputs for unseen data. Supervised learning is the most important methodology in machine learning and it also has a central importance in the processing of multimedia data. In supervised learning (often also called directed data mining) the variables under investigation can be split into two groups: explanatory variables and one (or more) dependent variables.

The target of the analysis is to specify a relationship between the explanatory variables and the dependent variable as it is done in regression analysis. To apply directed data mining techniques the values of the dependent variable must be known for a sufficiently large part of the data set. Supervised learning requires that the target variable is well defined and that a sufficient number of its values are given.

It is two steps process:

1. **Training step (Model Construction):** Learn classifier / regressor from training data. In this step, training data are provided to Classification Algorithm from which Classification model is constructed.
2. **Prediction step (Use the Model in Prediction):** Assign class labels/functional values to test data. In this step, Classifier prepared during 1st step is used to predict label for Unseen Data.

2.2 UNSUPERVISED LEARNING:

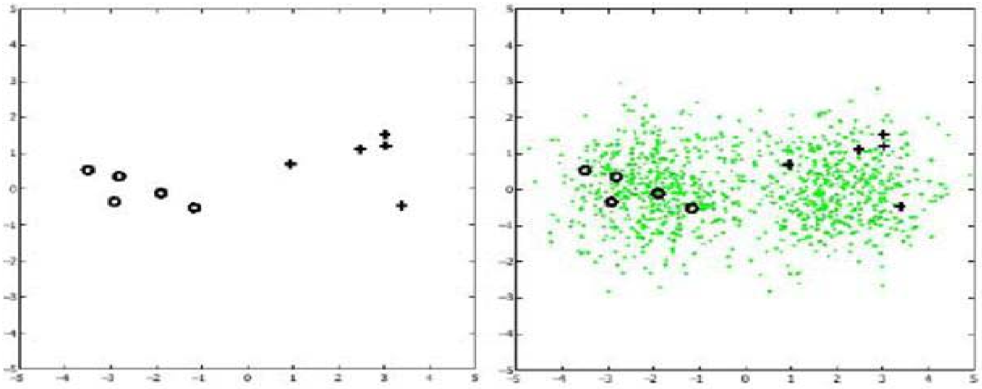
Unsupervised learning is learning in the absence of label or it can be stated as learning without training data. In unsupervised learning situations all variables are treated in the same way, there is no distinction between explanatory and dependent variables. However, in contrast to the name undirected data mining there is still some target to achieve. This target might be as general as data reduction or more specific like clustering. The dividing line between supervised learning and unsupervised learning is the same that distinguishes discriminate analysis from cluster analysis.

2.3 SEMI-SUPERVISED LEARNING [8]:

Semi-supervised learning (SSL) is a class of machine learning techniques that make use of both labeled and unlabeled data for training - typically a small amount of labeled data with a large amount of unlabeled data.

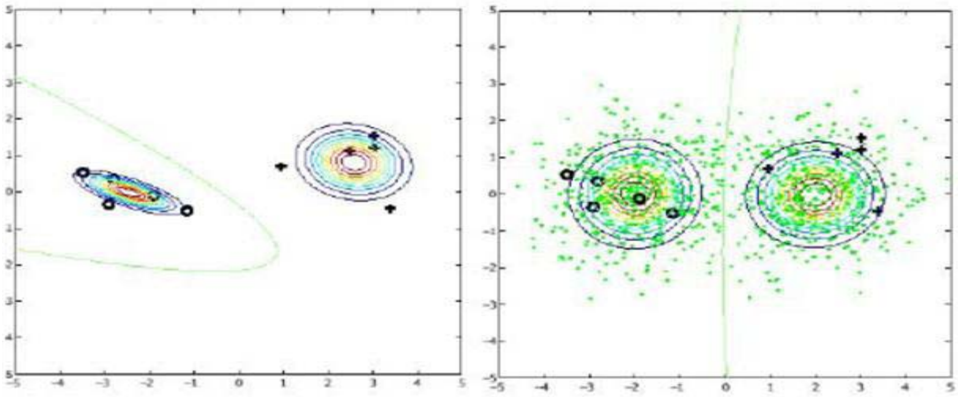
2.3.1 Importance of Semi-Supervised Learning: SSL is a special form of classification. Traditional classifiers use only labeled data (feature / label pairs) to train. Labeled instances however are often difficult, expensive, or time consuming to obtain, as they require the efforts of experienced human annotators. Meanwhile unlabeled data may be relatively easy to collect, but there has been few ways to use them. Semi-supervised learning addresses this problem by using large amount of unlabeled data, together with the labeled data, to build better classifiers. Because semi-supervised learning requires less human effort and gives higher accuracy, it is of great interest both in theory and in practice.

In the design of semi-supervised algorithms, the use of unlabeled data can be very useful. The main reason is that exploring unlabeled data may give rise to some evidence of the unknown distribution the examples are drawn from. Labeled examples are given and each class has a Gaussian distribution. Depicting a few examples from this Gaussian distribution, Figure 2.2 (a), it is difficult to discover the correct parameters of the Gaussian distributions, Figure 2.2 (c). However, using both, labeled and unlabeled data, Figure 2.2 (b), the parameters of the model can be discovered, Figure 2.2 (d). [4]



a) Labeled Data

b) Labeled and Unlabeled Data



c) Model learned from Labeled Data

d) Model learned from Labeled and Unlabeled Data

Figure 2.2: Use of unlabeled data to help parameter estimation in binary classification [4]

The iterative process in the standard EM-based semi-supervised learning includes two steps: firstly, use the classifier constructed in previous iteration to classify all unlabeled samples; then, train a new classifier based on the reconstructed training set, which is composed of labeled samples and all unlabeled samples. There is a problem in the process of reconstructing the training set, Some unlabeled samples are misclassified by the current classifier because the initial labeled samples are not enough, and the performance of the classifier is not well. These misclassified samples are considered directly as training samples, and used to construct a new classifier. This process affects the classification performance and convergence efficiency.

2.3 COMPARISION OF ALGORITHMS PROPOSED BY RESEARCHERS

Criteria	Reference Papers					
	[1]	[2]	[3]	[5]	[6]	[7]
Dataset Used	E-mails/ benchmark spam filtering corpora (PU1 & LINGSPAM)	Public spam e-mail dataset	Chinese Short documents of different categories.	TREC07p corpus	DARPA 1999 dataset	E-mails
Distribution of Dataset uniform	Yes	Yes	Yes	NS	NS	NS
Training, Testing Split	1 set for testing and 1 for training.	NS	NS	NS	NS	NS
Parameters compared for Accuracy	NS	Accuracy for diff algo on parameter AUC & F1	Time of iteration Vs. macro F1	True positive rate Vs. false positive rate	Classification accuracy Vs. number of labeled alerts	NS
Measures of evaluation used	Accuracy Measure f1	Accuracy Measure F1	Macro F1	ROC Curve	Classification Accuracy	AUC and Measure F
Method used for initial distribution of EM	NOT SSL	NOT SSL	NB	NB	NB	NS
Feature Selection method used	IG and TFV(term frequency variance)	NS	MI(Mutual Information)	TF-IDF	NS	Feature fusion
Uses more than one classifier	Yes	No	Yes	Yes	Yes	Yes

NS – Not specified

Table 1.1 shows comparison of all approaches of different papers discussed above by different criteria like, dataset used, distribution of dataset, training-testing split used, parameters compared for accuracy, method used for initial distribution of EM, feature selection method used , approach is using more than one classifier , what problem is addressed by authors in basic EM Algorithm etc.

In [1], they consider a task of threaten e-mail detection. E-mail classification is supervised learning problem. In this paper, they used the TCETHREATEN2 corpus with the standard bag of words representation and IG (Information Gain) for feature selection. Here they also used the TFV (Term Frequency Variance). They performed different algorithms like DT (Decision Tree), SVM (Support vector machine) and NB (Naïve Bayes). From this three algorithm DT outperforms in term of classification performance. DT is easy to tune & runs more efficiently on large dataset with high number of feature which makes it very attractive for text classification. And they found that the feature selector IG performs better than the TFV.

As per [3], in the iterative process of EM, reconstructing the labeled training samples is taken into account. Because the labeled samples are limited and the performance of the classifier is not well, the labels of some unlabeled samples are not confidently, which are derived by the classifier constructed based on the labeled samples. If these misclassified samples are incorporated into the labeled training samples and then considered as a part of reconstructed labeled training set to train a new classifier, they will disrupt the normal process of learning and reduce the classification performance to some extent. On the other hand, some samples are easy to be classified correctly in the current classifier. In order to enrich the information of current classifier, these reliably samples should be added to the labeled training set as soon as possible. Meanwhile, these reliable unlabeled samples are considered as labeled samples and retain in the next iteration, which is beneficial to reduce the amount of unlabeled samples.

In [5], such classifiers often require a large training set of labeled emails to attain a good discriminant capability between spam and legitimate emails. In addition, they must be frequently updated because of the changes introduced by spammers to their emails to evade spam filters. To address this issue active learning and semi-supervised learning techniques can be used. However, users are usually willing to label only a few emails, and the benefits of self-training techniques are limited. In this paper they propose an active semi-supervised learning method to better exploit unlabeled emails, which can be easily implemented as a plug-in in real spam filters. This method is based on clustering unlabeled emails, querying the label of one email per cluster, and propagating such label to the most similar emails of the same cluster. The effectiveness of our method is evaluated using the well-known open source Spam Assassin filter, on a large and publicly available corpus of real legitimate and spam emails.

In [6], they give a semi-supervised alert classification model which makes use of the power of semi supervised learning. Here, they have used EM based algorithm for classification and Method used for initial distribution of EM is Naïve Bayes. For supervised learning, it may need long time and expertise of network security to manually label the alert data. As per their review, by using alert context properties, accuracy is increased by about 3 percent.

In [7], they describe a machine learning approach for detecting web spam. Their approach involves adding human-engineered features and then using semi-supervised learning to exploit the unlabeled data that are provided as part of the web spam challenge data. They also use their combinational feature-fusion approach in order to reduce the number of TF-IDF content based features and to construct new features that are combination of these features. They have implemented ADT (Alternative decision tree), SVM (support vector machine), and NB (naïve bayes). From these algorithms ADT gives better performance than other algorithm. They have implemented these algorithms with three different manners: 1) Initially 2) Semi-supervised 3) Semi-supervised + feature-fusion. After comparing the results of these three manners, 3rd method gives better performance than the others. The limitation of this paper is that since they assign a rank based only on positive and negative magnitude of the score value, their method will not handle the case well where the most predictive value occurs in middle. And as future work, they try to generate more sophisticated link based feature how the number of iteration of semi-supervised learning impacts classifier performance.

CONCLUSION

Semi-Supervised Learning can be effectively used for improving performance of Classification when limited numbers of labeled documents are available for training. To solve problem of misclassified unlabeled documents added in labeled documents set of in each iteration of classic algorithm which include ensemble learning using k-NN and NB to include only reliable labeled documents to training set in each iteration to increase accuracy.

REFERENCES:

[1] S. Appavu and R.Rajaram, "*Learning to classifying threaten email*", 2008 IEEE.

[2] Lei SHI, Qiang WANG "Spam *e-mail classification using Decesion tree Ensemble*", 2012.

[3] Xinghua Fan and Houfeng Ma, "An *improved EM-based Semi-supervised learning method*", 2009 IEEE.

[4] Xiaojin Zhu, "Semi-*Supervised Learning Literature Survey*", Computer Sciences TR 1530, University of Wisconsin – Madison, 2005.

[5] Jun-ming Xu, Giorgio Fumera, Fabio Roli and Zhi-Hua Zhou "*Training SpamAssassin with Active Semi-supervised Learning*", CEAS 2009.

[6] Haibin Mei and Minghua zhang, "*A semi supervised IDS alert classification model based on alert context*", ICCSEE 2013.

[7] Ye Tian, Gary M. Weiss and Qiang Ma, "*A semi-supervised approach for web spam detection using combinatorial feature-fusion*", 2007.

[8] Xiaojin Zhu, "Semi-*Supervised Learning Literature Survey*", Computer Sciences TR 1530, University of Wisconsin – Madison, 2005.

[9] Vinod Patidar, Divakar Singh, "*A Survey on Machine Learning Methods in Spam Filtering*", International Journal of Advanced Research in Computer Science and Software Engineering, Page(s): 964-972, October 2013

[10] MohammadReza FeiziDerakhshi and Nayer TalebiBeyrami, "*The Feature Selection and Dimensionality Reduction Methods for Email Classification*", Journal of Basic and Applied Scientific Research , 633-636, 2013