International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249–8958, Volume-1 Issue-3, February 2012

A Survey and Comparative analysis of Expectation Maximization based Semi-Supervised Text Classification

Purvi Rekh, Amit Thakkar, Amit Ganatra

Abstract*: *Semi-supervised learning (SSL) based on Naïve Bayesian (NB) and Expectation Maximization (EM) combines small limited numbers of labeled data with a large amount of unlabeled data to help train classifier and increase classification accuracy. 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 are limitations of standard EM-based semi-supervised learning like, problem in the process of reconstructing the training set - some unlabeled samples are misclassified by the current classifier, problem of over-training, problem of as the number of documents increases, the running time increases significantly. With the aim of improving the efficiency problem of the standard EM algorithm, many authors have proposed approaches. These approaches are described in this paper, also comparison of these approaches is done and limitations of these methods are described. Also some research challenges are given in this area.

Index Terms*: *Expectation Maximization, Naïve Bayesian, Semi-supervised learning, Text Classification.

I. INTRODUCTION

Consider the problem of automatically classifying text documents. In many text domains, especially those involving online sources, collecting unlabeled documents is easy and inexpensive, but labeling such documents is expensive and requires experts. Accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents [2].

Semi-supervised learning is a class of machine learning techniques that make use of both labeled and unlabeled data for training. Semi-supervised learning is one of the mainstream methods for exploiting unlabeled examples in addition to labeled ones to improve learning performance.

Manuscript published on 28 February 2012.
* Correspondence Author (s)
Purvi K. Rekh*, U & PU Patel department of computer engineering, Chandubhai S Patel institute of technology, Changa, Petlad, India, 9909793291, (email: purvirekh@gmail.com).
Amit R. Thakkar, Department of Information and Technology, Chandubhai S Patel institute of technology, Changa, Petlad, India, 9601290990, (e-mail: amitthakkar.it@ecchanga.ac.in).
Amit Ganatra, U & PU Patel department of computer engineering, Chandubhai S Patel institute of technology, Changa, Petlad, India, 9426350639, (email: amitganatra.ce@ecchanga.ac.in).

Semi-supervised learning algorithms can be divided into four categories: generative methods, graph-based methods, Co-training methods, and S3VMs (Semi-Supervised Support Vector Machines). Among them, generative methods have been widely applied to text classification. In [2], Nigam et al. introduced an algorithm for learning from labeled and unlabeled documents, based on the combinations of the EM (expectation maximization) algorithm with naïve Bayesian classifiers, and show that the accuracy of learned classifiers be improved by using large amount of unlabeled documents together with the labeled training ones.

The algorithm first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. It then trains a new classifier using the labels for all the documents, and iterates to convergence. This basic EM procedure works well when the data conform to the generative assumptions of the model. However these assumptions are often violated in practice, and poor performance can result.

To overcome this limitation, several researchers have presented improvements in classic EM algorithm to improve performance which are described in this paper.

The rest of this paper is organized as follows. In Section II, Comparison of methods used for Semi-Supervised Learning is shown as well as merits and demerits of different Semi-Supervised Learning methods are described. In Section III, First, how Naïve Bayes classifies Text document is described and then approaches that use Naïve Bayes for finding initial classifier for EM are described and then approach that use Random Sub-Space method as initial step for EM is described. In Section IV, comparison of all approaches described in section III is shown. Finally in section V and VI, research challenges and Conclusion are given respectively.

II. METHODS OF SEMI-SUPERVISED LEARNING

Table I shows different approaches, algorithm used for that approach and assumption that is made for using that approach [7]. Table II shows merit and demerit of all approaches [7].

III. RELATED WORK

Among the approaches shown in Table I,

this paper focuses
on generative method which make use of Naïve Bayes and Expectation Maximization. So in this section, first, how Naïve Bayes classifies Text document is described and then approaches that use Naïve Bayes for finding initial classifier for EM are described. Then how Random Subspace Method classifies Text documents is described and approach that uses Random Subspace Method for finding initial classifier for EM is described.

Table I: Methods of Semi-Supervised Learning

Approach	**Example Algorithm / Method**	**Assumption**
Low Density Separation	Semi-Supervised Support Vector Machine $(S^3VM.)$	Cluster Assumption
Graph based Methods	Build weighted graph (w_{kl}) $\min_{(y_j)} \sum_k \sum_l w_{kl}(y_k - y_l)^2$	Manifold Assumption
Co –Training	Train two predictors $y_j^{(1)}$, $y_j^{(2)}$ Couple objectives by adding $\sum_j (y_j^{(1)} - y_j^{(2)})^2$	Independent Views
Generative Method	Naïve Bayesian Expectation Maximization	Rely on a model for the distribution of the input data

Table II: Merits and Demerits of methods of Semi-Supervised Learning

Methods	**Merits**	**Demerits**
S^3VM	→Applicable wherever SVMs are applicable →Clear mathematical framework	→Optimization difficult →Can be trapped in bad local optima →More modest assumption than generative model or graph-based methods, potentially lesser gain
Graph based Methods	→Clear mathematical framework →Performance is strong if the graph happens to fit the task →The (pseudo) inverse of the Laplacian can be viewed as a kernel matrix →Can be extended to directed graphs	→Performance is bad if the graph is bad →Sensitive to graph structure and edge weights.
Co –Training	→Simple wrapper method. Applies to almost all existing classifiers	→Natural feature splits may not exist →Models using both features should do better

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Generative Method	→Clear, well-studied probabilistic framework →Can be extremely effective, if the model is close to correct	→Often difficult to verify the correctness of the model →Model identifiability →EM local optima →Unlabeled data may hurt if generative model is wrong

Table III: Classic Semi-Supervised Algorithm

Inputs: Collections D^l of labeled documents and D^u of unlabeled documents.
Method :

- Build an initial naive Bayes classifier, $\hat{\theta}$, from the labeled documents, D^l, only. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \text{argmax}_{\theta} P(D \mid \theta)P(\theta)$
  (Equation 5.6 in [2])
- Loop while classifier parameters improve, as measured by the change in
  $l_C(\theta \mid D, z)$
  (the complete log probability of the labeled and unlabeled data, and the prior) (Equation 10 in [2]):
  - **(E-step)** Use the current classifier, $\hat{\theta}$, to estimate component membership of each unlabeled document, *i.e.*, the probability that each mixture component (and class) generated each document, $P(c_j \mid d_i; \hat{\theta})$ (Equation 7 in [2]).
  - **(M-step)** Re-estimate the classifier, $\hat{\theta}$, given the estimated component membership of each document. Use maximum a posteriori parameter estimation to find $\hat{\theta} = \text{argmax}_{\theta} P(D \mid \theta)P(\theta)$.

Output: A classifier, $\hat{\theta}$, that takes an unlabeled document and predicts a class label.

A. *Text Classification using Naïve Bayesian*

The algorithm in Table III, first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents. The naive Bayesian technique is a popular method which probabilistically labels the unlabeled documents for text categorization. Let D= {d_1, d2, ..., d_n} be the training document set and C = {C1, C2, . , C_n} be a set of predefined classes. Each document can be represented as an ordered list of words. The vocabulary, V= {w_1, w_2, …w_n}, is the set of all words considered for classification. The Naive Bayes classifiers estimate the posterior probability $P(c_j \mid d_i)$ which represents the probability that a document d_i belongs to a class Cj. Using the Bayes rule, we have:

$$P(c_j \mid d_i; \hat{\theta}) = \frac{P(c_j \mid \hat{\theta}) \prod_{k=1}^{|d_t|} P(w_{d_{t,k}} \mid c_j; \hat{\theta})}{\sum_{r=1}^{|C|} P(c_r \mid \hat{\theta}) \prod_{k=1}^{|d_t|} P(w_{d_{t,k}} \mid c_r; \hat{\theta})}$$

Maximum a posteriori parameter estimation is performed by:

$$P(w_t \mid c_j; \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} N(w_t, d_i) P(y_i = c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(w_s, d_i) P(y_i = c_j \mid d_i)}$$

The class same way, the prior probability parameters are set by the maximum likelihood estimate:

$$P(c_j \mid \hat{\theta}) = \frac{1 + \sum_{i=1}^{|D|} P(y_i = c_j \mid d_i)}{|C| + |D|}$$

where $N(w_s,d)$ is the frequency of word w_s occurs in document dj, and $P(c_j|d_i) \in \{0, 1\}$.

B. *Discussion of approaches by different authors that uses Naïve Bayesian as initial classifier.*

In [2], Kamal Nigam introduced an algorithm for learning from labeled and unlabeled documents based on the combination of Expectation-Maximization (EM) and a Naive Bayes classifier. The algorithm first trains a classifier using the available labeled documents, and probabilistically labels the unlabeled documents.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

It then trains a new classifier using the labels for all the documents, and iterates to convergence.

This basic EM procedure works well when the data conform to the generative assumptions of the model. This paper shows that the accuracy of learned text classifiers can be improved by augmenting a small number of labeled training documents with a large pool of unlabeled documents. Standard (classic) Semi-supervised EM algorithm is given in table III.

K. Nigam addressed the problem that basic EM algorithm can suffer from a misfit between the modeling assumptions and the unlabeled data. To solve this problem, he provided two solutions:

1) Introduces a weighting factor that dynamically adjusts the strength of the unlabeled data's contribution to parameter estimation in EM.

2) Reduces the bias of naive Bayes by modeling each class with multiple mixture components, instead of a single component.

In [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 [4], authors considered the same problem that the classification performance is not well when the count of the initial labeled samples is very small and provided solution for limitations of solutions provided for the same problem in [3]. These limitations are:

1) Division mechanism is not clearly described.

2) The impact of incorrect samples in the reliable training set is still not considered.

Authors proposed a semi-supervised method based on incremental EM algorithm. This method makes full use of the useful information of intermediate classifier. On the one hand, this method verifies the feasibility of division existed in unlabeled samples, and uses the division mechanism to enhance the reliability of new incremental samples by dividing the unlabeled samples scientifically; on the other hand, a feedback learning mechanism is proposed, and it is used to decrease the probability of adding misclassified samples.

In [5], authors addressed 2 problems: First is main difficulty of automatic text classification, that the dimensions of the feature space are tremendous. It usually reaches to thousands and even hundreds of thousands. Thus the feature selection (FS) becomes a crucial step in classification. The class unbalance is ubiquitous in text categorization. It increases the complexity and difficulty for classification. Another problem addressed is that, though semi-supervised classification with the basic EM obtains good achievements, it has unavoidable problem of classification efficiency. As the number of documents increases, the running time increases significantly.

Authors have provided solutions for these problems: Firstly, a feature selection function of strong category information is constructed to control the dimension of feature vector and preserve useful feature terms. Secondly, an intermediate classifier gradually transfers unlabeled documents of maximum posterior category probability to labeled collection during each iteration process of the EM algorithm. The iteration number of the enhanced EM is less than the basic EM. Finally, experiments shows that the improved method obtains very effective performance in terms of macro average accuracy and algorithm efficiency.

In this paper, an effective feature selection function is constructed to filter a large number of invalid feature words and keep high categories information words down. At the same time, the EM algorithm with Naïve Bayesian is also adjusted by transferring unlabeled document possessing maximum confidence from unlabeled set to labeled set in each step. The experiment results indicate that learning velocity and macro average accuracy of the enhanced EM are better than the basic algorithm.

In previous discussions, all authors have used NB as an initial classifier for EM. In [6], Authors have used Random Subspace Method as an initial classifier for EM.

C. Random Subspace Method [6]

The stochastic discrimination (SD) theory constructs an ensemble classifier by many stochastically created weak component classifiers in order to achieve accuracies higher than those obtained from a single classifier. SD is characterized by the properties of overtraining-resistance, a high convergence rate, and a low misclassification error rate. Random Subspace Method (RSM), introduced by Ho, is one of the stochastic discrimination (SD) methods based on a stochastic feature space sub sampling process. RSM is a very simple and popular ensemble construction method.

In [6], authors incorporate RSM into in the framework of the EM algorithm to improve the classification performance and avoid to over-training. The random subspace approach randomly samples a subset of features from the entire set of features space and then constructs a classifier on each random set of features, and combines them using a heuristic such as majority voting, sum rule, etc as shown in Figure 1. Specifically, suppose that the dimension of original feature space is n, the dimension of random subspaces is m, m<n. In the RSM, by projecting all the data in the n-dimensional training set onto the m-dimensional subspace, we get the m-dimensional random subspace. This is repeated K times to build K different views of the feature space which are then used to train K base classifiers. In order to improve the learning performance, authors combine random subspace method with the basic EM algorithm which is called RS-EM. RS-EM subspaces of the feature space, and trains each of the subspaces to assure the random classifier on subspaces is constructing different

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

classifiers. RS-EM uses ensemble classifier to predict the labels of the unlabeled data, and choose the most confident data to enlarge the training data set of the classifiers. The combination of K classifiers is computed by:

$\theta = \frac{1}{K} \sum_{i=1}^{K} \theta i$

Then, authors choose the probabilistically assigned labels into the newly labeled data set .The above process is repeated for some number of times. In each of the iteration, the K classifiers are trained with enlarged labeled data set. In this way, unlabeled data is used to boost the performance of standard supervised learning.

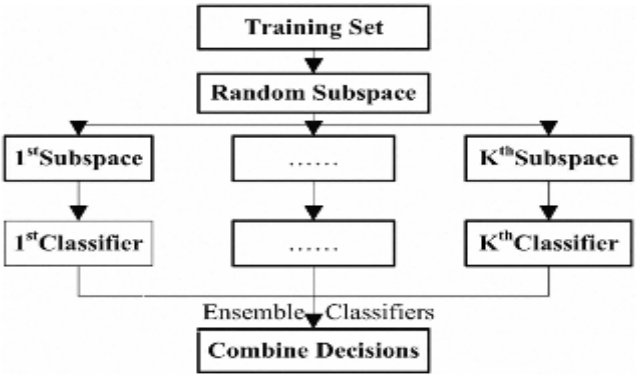

Figure 1 The framework of the RSM

Table IV: Comparison of proposed approaches in reference papers by different criteria

Criteria	Reference papers				
	[2]	[3]	[4]	[6]	[5]
Dataset Used	1) 20 News Group 2) WebKb 3)Reuters	Chinese Text	Chinese Short Text	Text document from public forums in Chinese internet	Reuter 21578
Distribution of Dataset uniform	NS	Yes	Yes	NS	No
Training, Testing Split	NS	NS	NS	3/4, 1/4	2/3, 1/3
Parameters compared for Accuracy	No of Unlabeled Documents vs. Accuracy	Times of iteration vs. Macro F1	Times of iterations vs. Macro F1	No of iterations vs. Accuracy	Feature Selection methods vs. Accuracy
Measures of evaluation used	Accuracy	1) Macro F1 2)New measure, IR = (IS – IL)/IL	NS	Macro F1	Macro average Accuracy
Method used for initial distribution of EM	Naïve Bayesian	Naïve Bayesian	Naïve Bayesian	Random Sub-Space method	Naïve Bayesian
Feature Selection method used	NS	TF-IDF in each iteration	Chi-Square in each iteration	NS	DF * ICIF
Uses more than one classifier	No	Yes	No	Yes	No
Problem Addressed in Basic Algorithm	The basic EM algorithm can suffer from a misfit between the modeling assumptions and the unlabeled data.	Reconstructing the training samples. Some unlabeled samples are misclassified by the current classifier.	Reconstructing the training samples. Some unlabeled samples are misclassified by the current classifier.	Over-training	As the number of documents increases, the running time increases significantly.
Improvement in accuracy / other improvement	Improvement in accuracy after adding unlabeled documents.	Improvement in F1 and IR	Improvement in F1	4.54% compared to Semi-supervised EM and 8.16% compared to naive Bayes	1) Improvement in Micro average accuracy. 2) Number of iterations are less than the basic EM.

*NS – Not Specified

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

IV. COMPARISON OF ALL APPROACHES

Table IV 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.

V. RESEARCH CHALLENGES

Research challenges in this area are as follow:

1) Feature selection and text pre-processing methods have lots of impact on accuracy of text classification. So one of the research challenges include to test effect of various pre-processing and feature selection methods on Semi-Supervised Learning [9].

2) Also in [3] and [4], the mechanism of division should be improved.

3) In [5], the choosing of some parameters (e.g. the threshold of feature selection function DF*ICIF, and the number of labeled documents etc.) are needed to be further investigated.

4) Also to test the effect of Semi-Supervised Learning on different datasets like poem and sentimental analysis is also a research challenge [10].

VI. CONCLUSION

Semi-Supervised Learning with EM can be effectively used for improving performance of Text Classification when limited numbers of labeled documents are available for training. Many authors have suggested solutions for problems like misfit between the modeling assumptions and the unlabeled data, problem of over training, problem in the process of reconstructing the training samples, problem of as the number of documents increases, the running time increases. But there is no universal method to solve all the problems and still improving accuracy and reducing training time of text classification using Semi-Supervised Learning is an issue.

REFERENCES

[1] Vishal Gupta, "A Survey of Text Mining Techniques and applications", Journal of Emerging Technologies, In Web Intelligence, Vol. 1, No. 1, August 2009.

[2] Kamal Nigam, Andrew Kachites Mccallum, " Text classification from Labeled and Unlabeled Data using EM", Machine Learning, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 2002 H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985, ch. 4.

[3] Xinghua Fan, Zhiyi Guo, Houfeng Ma. "An improved EM-based Semi-supervised Learning Method" ,International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, page(s): 529 - 532, August – 2009.

[4] Xinghua Fan, Zhiyi Guo; "A semi-supervised Text Classification Method based on Incremental EM Algorithm", WASE International Conference on Information Engineering, Page(s): 211 - 214, 2010.

[5] Wen Han, Xiao Nan-feng, "An Enhanced EM Method of Semi-supervised Classification Based on Naive Bayesian", Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), 15- Sep- 2011.M. Young, *The Techincal Writers Handbook.* Mill Valley, CA: University Science, 1989.

[6] YueHong Cai, Qian Zhu; "Semi-Supervised Short Text Categorization based on Random Subspace"- Computer Science and Information Technology (ICCSIT), 3rd IEEE International Conference on Page(s): 470 – 473 , 2010.J. Jones. (1991, May 10). Networks (2nd ed.) [Online]. Available: http://www.atm.com

[7] Xiaojin Zhu, "Semi-Supervised Learning Literature Survey", Computer Sciences TR 1530, University of Wisconsin – Madison, 2005.E. H. Miller, "A note on reflector arrays (Periodical style—Accepted for publication)," *IEEE Trans. Antennas Propagat.*, to be published.

[8] Yutaka Sasaki, "Automatic Text Classification", NaCTeM, School of Computer Science.C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[9] GuoQiang, "Research and Improvement for Feature Selection on Naïve Bayes Text Classifier", 2nd International Conference on Future Computer and Communication, Volume 2.

[10] Bei Yu, "Evaluation of Text classification methods for literature survey", Literary and Linguistic Computing, Vol. 23, No. 3, 2008.

AUTHOR PROFILE

Purvi Rekh has received her B.E degree in Information Technology from Veer Narmad South Gujarat University, Gujarat, India in 2007 and pursuing master Degree from Charotar University of science and Technology, Changa. She has worked as an adhoc lecturer in Sarvajanik College of Engineering and Technology, Surat, Gujarat from 2008 to 2010. Her current research interest is Semi-Supervised Text Classification.

Amit Thakkar has received his B.E degree in Information Technology from Gujarat University, Gujarat, India in 2002 and master Degree from Dharmsinh Desai University, Gujarat, India in 2007. He has joined his Ph.D in the area of Multi relational Classification at Kadi Sarvavishvidhalaya University, Gandhinagar, India in June 2010.

Since 2002 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Information Technology. He has published more than 20 research papers in the field of data mining and web technology. His current research interest includes Multi relational Data Mining, Relational Classification and Associate Classification.

Amit Ganatra has received his B.E degree in Computer Engineering from Gujarat University, Gujarat, India in 2000 and master Degree from Dharmsinh Desai University, Gujarat, India in 2004. He has joined his Ph.D in the area of Multiple Classifier System (Information Fusion) at Kadi Sarvavishvidhalaya University, Gandhinagar, India in August 2008.

Since 2000 he has been with faculty of Engineering and Technology, Charotar University of Science and Technology, Changa, Gujarat, Where he is currently working as an Associate Professor in the Department of Computer Engineering. He has published more than 50 research papers in the field of data mining and Artificial Intelligence. His current research interest includes Multiple Classifier System, Sequence Pattern Mining.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)