Active Learning for Reducing Bias and Variance of a Classifier Using Jensen-Shannon Divergence

Minoo Aminian
Computer Science Dept.
SUNY Albany, Albany
NY, USA, 12222
minoo@cs.albany.edu

Abstract

We consider reducing loss of a classifier by decreasing its bias and variance. Embarking upon classification of scarcely labeled data, we use active learning approach in semi-supervised learning, and show that we can speed up convergence to a desired level of loss. Our focus, in this paper, is on the best instance selection for labeling the unlabeled data; we use Jensen-Shannon divergence as one selection criterion. We show that our single instance selection approaches are superior to multiple selection approach. Empirical results indicate that this method can decrease classification loss significantly.

1. Introduction

Learning from labeled and unlabeled data is a complex domain. Reducing classification loss, when labels are difficult to obtain, has always posed a challenge; selecting the best unlabeled instances to label has posed another challenge. We propose an active learning method aimed to meet both challenges. This method reduces the loss of the classifier via reducing its bias and variance, and speeds up the convergence of the learner to certain level of loss by selecting the most informative instances.

Solving the problem of semi-supervised learning basically relies on one of the two existing approaches: passive and active semi-supervised learning. In passive semi-supervised learning [19], [20], we can train a classifier based on the labeled data as well as the unlabeled data. Typically, the labels for unlabeled data are imputed by certain means based on the current state of the classifier. The now augmented labeled data is then used to retrain the classifier [21].

In active learning first we train a classifier from the labeled data. Then we select the instances from the unlabeled data which if labeled and added to the labeled data are likely to greatly enhance the performance of the classifier. Next we ask an oracle or a human to label these instances and add them to the set of labeled data to retrain the classifier. This process can be repeated with active learning aim of labeling as little number of data as possible to reach certain level of accuracy.

The active learning work of Cohen [6] describes a method that attempts to minimize the error of a *regression* learner by minimizing its estimated squared bias. This method selects actions/queries designed to minimize the bias of a locally weighted regression-based learner. He applies his method on two simple regression problems in which the solution for the expected bias can be found in closed form, but unfortunately for many tasks and models this optimal selection can not be found in closed form.

Sometimes we can make use of local solutions to compensate for non-optimal search. For example, another method called co-training [1] is a semi-supervised algorithm applicable to problems with two separate but redundant views of the data. Starting with a small set of labeled data and a large pool of unlabeled data, co-training bootstraps the views from each other to boost the accuracy of the initial classifier.

Multi-view co-training algorithms, though successful, rely on the ability to factor the available features into two independent and compatible views. Therefore, they are not applicable to problems with no obvious feature split. As a result, researchers have begun to investigate the co-training procedures that use two different learning algorithms in lieu of the multiple views required by standard co-training [4].

For the scarcely labeled datasets, the setting of this paper, the learner has a relatively small amount of labeled data and a large number of unlabeled data. It is assumed that noise is uniform and both labeled and unlabeled data are drawn from the same population, but the process of labeling is too costly or difficult. We propose a single-view algorithm for bootstrapping two classifiers from the labeled data, one a single naïve Bayes and the other by applying bagging on the initial labeled data and treating the result as the prediction of a single classifier. Bagging can be viewed as an approximation of Bayesian model

0-7695-2495-8/05 $20.00 © 2005 IEEE

averaging by importance sampling [17]. Bayesian optimal classifier can be defined to give on the average the most probable prediction given the training data but since the labeled data in situations which we study is scarce neither of our two initial classifiers alone is good enough to classify the whole data. Therefore, we use two different learning algorithms to obtain two different views and learn from their differences. We show that our method faced with scarcely labeled data through active learning can outperform both initial classifiers including the second classifier which is an approximation to Bayesian optimal classifier. We also propose improvement to our algorithm based on special instance selection policy using Jensen-Shannon divergence and show that we can achieve even higher reduction in loss using this strategy.

Jensen-Shannon divergence [23] is a new measure based on Jensen's inequality and Shannon entropy. It is a useful measure of the distance between distributions and can be used to evaluate the potential utility of instances [22].

As number of labeled data increases we can either continue using this model or employ another model applicable to partially labeled data. For the scope of this paper, we will continue using this method; but in future work we will compare its performance with other methods.

We begin the paper by giving the notation and settings; next we present our algorithm, discuss instance selection criteria, and explain how we pick the most informative instances to reduce bias and variance of the learner via active learning. We demonstrate the empirical results to verify the idea, and finally discuss the associated issues and future work.

2. Notation and background

Though the bias variance decomposition in its original formulation for squared error loss [7] is useful, it is not readily applicable to *classification* problems with 0-1 loss function. There are several corresponding decompositions for zero-one loss, in this paper, we use the bias-variance decomposition proposed by Domingos [2].

Suppose we have a training set of pairs $\{(x_i, t_i), i = 1,...,n\}$, and a model which produces an estimate y_i for x_i. Let t be the true value (most probable value) of the estimated variable for the test example x. It is important to realize that Domingos treats the instance labeling using a uniform noise model, hence there is a chance of mislabeling of an instance. The zero-one loss is zero if $y = t$, and is one otherwise. There are many definitions to bias and variance in classification, for example to understand Domingos definition, we need to define the notions of optimal prediction and main prediction. The optimal prediction for a specific example x is the lowest loss prediction irrespective of our model or formally:

$$y_* = \arg\min_{y_i} E_t[L(t, y_i)] \qquad (1)$$

And the main prediction, y_m for the specific value x a specific loss function L and a set of training sets D, is defined to be the value that differs least from all other predictions y according to L.

$$y_m = \arg\min_{y'} E_D[L(y_i, y')] \qquad (2)$$

Then bias of a learner on a specific example is defined as:

$$B(x_i) = L(y_*, y_m) \qquad (3)$$

And the variance of the learner on an example as:

$$V(x_i) = E_D[L(y_i, y_m)] \qquad (4)$$

and based on all these the following decomposition holds:

$$E_{D,t}[L(t,y_i)] = c_1N(x_i) + B(x_i) + c_2V(x_i) \qquad (5)$$

In which c_1 and c_2 are multiplicative factors which will take different values for different loss functions.

So, assuming a uniform noise, we propose a method that, based on equations (3) and (4), reduces the expected value of bias and variance of the learner through selection of specific instances via active learning. We will explain this method in the following section.

3. Active learning algorithms

Given a set of scarcely labeled data, we start our algorithm by training the initial classifier, h_1. This classifier is a naïve Bayes classifier trained on the labeled data, but since the amount of labeled data is small, does not provide high accuracy.

Next, we apply bagging with naïve Bayes as the underlying classifier to the labeled data, and return the class that has been predicted most often. We treat this result as a single second learner, h_2. Then we pick a random sample pool, R, from the unlabeled data, apply both h_1 and h_2 to R, and compare their predictions. Wherever they disagree on labeling an instance(s) we either give all those instances (MIS section 3.1) to the oracle for labeling or select one instance (SIS section 3.2 or SIS-JSD section 4.2) to give to the oracle to provide the correct label(s), add the instance(s) to the labeled set and retrain both classifiers. Again we take another random pool from the unlabeled data and apply classifiers h_1 and h_2 to this pool, compare the predictions of the two learners, and ask the oracle to label the ones they disagree upon. We continue this process and propose that repeating this process in a loop for 0-1 loss, reduces the loss of the learner. To show this phenomenon we examine what happens in each iteration. For the simplicity we consider the two class problem first.

During each iteration, after comparing predictions of h_1 and h_2, whenever there is a disagreement on labeling an instance at least one of the predictions is wrong. The oracle, then gives the correct label to the instance. Therefore, for that specific instance h_1 has the optimal prediction y_* , (or $y = y_*$), and h_2 provides the majority voting, y_m. Now For the training sets that $y = y_*$, one of the following two cases will happen - case one: $y = y_m$ which in that case based on equations (3 and 4) the prediction for that example is unbiased and the variance is reduced causing reduction of overall bias, variance and loss. Case two: $y \neq y_m$ which in that case we can prove that for the training sets that $y = y_*$, but $y \neq y_m$ the variance contributes to reduction of loss in those training sets causing the overall loss to be reduced. This reasoning is based on theorem 1 and 2 in [2] which state that equation (5) is valid for zero-one loss in two class problems, with $c_1 = 2P_D(y = y_*) - 1$ and $c_2 = 1$ if $y_m = y_*$, $c_2 = -1$ otherwise. While in multi-class problems equation (6) is valid, with $c_1 = P_D(y = y_*) - P_D(y \neq y_*)P_t(y = t \mid y_* \neq t)$ and $c_2 = 1$ if $y_m = y_*$, $c_2 = -P_D(y = y_* \mid y \neq y_m)$ otherwise.

In other words, in multi-class problems for all the training sets for which $y \neq y_m$ if $y = y_*$, the variance contributes to reducing loss in those training sets and therefore reduces overall loss. Now for all the training sets that $y \neq y_m$ and $y \neq y_*$, providing the correct label by the oracle will reduce the bias which contributes to reduction of loss.

Thus, by selecting the specific instances with high bias and variance and correcting their labels, we either have reduced both bias and variance or only variance of the learner causing the average loss to reduce. Based on all these we can state the generalized version of our algorithm as following:

```
Given:
   a set $D_L$ of labeled training instances
   a set $D_U$ of unlabeled instances
Loop for k iterations to reach user satisfaction:
   • Train classifier $h_1$ from $D_L$.
   • Train classifier $h_2$ from $D_L$ using bagging.
   • Pick a random sample pool R from $D_U$.
   • Apply $h_1$ and $h_2$ to $R$ and compare their
     predictions.
   Let S be the set of instances from R on which $h_1$
    makes the predictions such that
        $\forall x \in S$ : $h_1(x) \neq h_2(x)$
   • $\forall x \in S$ Select x based on the selection strategy,
     label $x$ with the oracle and add it to the training
     set $D_L$.
```

Alg.1: Active learning to reduce bias and variance

3.1. Multiple instance selection algorithm MIS

In MIS selection strategy, we select all the instances that the two classifiers disagree on their labels to give to the oracle. The result of applying this selection strategy on a number of benchmark datasets from UCI repository is recorded in Table 1 and Figures 1, 2, and 3.

3.2. Single instance selection algorithm SIS

To improve our initial approach, considering the cost of using the oracle to label an instance, we propose the following selection strategy. Among the instances that the two classifiers disagree on their labels, in each iteration, we seek out the best instance to give to the oracle for labeling and propose that if we proceed with one selective instance at a time, the classifier will reach certain level of loss with less number of labeled data than if we proceed with more than one of those instances. The classifier following training by one more labeled instance can become more competent than before, hence better able to label. Selection of the best instance can depend on the specific case or given constraint(s), for example in the interactive labeling system in mature stages of the examples collection process, one can use the already collected examples to build a good classifier, which in turn can aide in the selection of examples [18]. For this algorithm, we select the instance in which both classifiers have the least confidence. Then, after having oracle label this instance we add it to the training set, retrain the classifiers, and repeat the process. The results of applying this selection strategy on dataset Pima Indians Diabetics is shown in Table 2 and Figure 4.

4. The Jensen-Shannon divergence metric

Jensen- Shannon (JS) divergence [11] is a useful measure of the distance between probability distributions particularly in the study of decision problems. The JS-divergence between the two probability distributions P_1 and P_2 is defined as:

$$\begin{aligned} &\mathrm{JS}_{\pi}(P_1, P_2) = \\ &H(\pi_1 P_1 + \pi_2 P_2) - \pi_1 H(P_1) - \pi_2 H(P_2) \end{aligned} \tag{6}$$

where π_1, $\pi_2 > 0$ are the weights assigned to P_i with the assumption that $\pi_1 + \pi_2 = 1$, and when the weights are equal JS_π is simply JS. H(P) is the Shannon entropy of the distribution P= $\{p_j : j=1,2,\ldots,k\}$ defined as:

$$H(P) = -\sum_{j=1}^{k} P_j \log P_j \tag{7}$$

We can use this measure here as a metric to measure the similarity between our two learners. If $P_i(x)$ is the class probability distribution given by the i^{th} classifier for the example x, we can compute the JS-divergence of the two classifiers that we have, and use that measure to select a specific example. The higher the value of JS-divergence, the greater spread in the predicted class probability, and a zero value for the JS-divergence indicates that the distributions are identical.

4.1. Single instance selection using JS-divergence

For this selection strategy, we compute the JS-divergence for the instances on which the two classifiers disagree upon their labeling and select the instance with the highest JS-divergence. Then, we have the oracle label this instance, add it to the training set, and retrain the classifiers. Repeating this process in each iteration will reduce the overall loss. Algorithm 2 presents the JS-divergence selection method.

```
Given:
  a set $D_L$ of labeled training instances
  a set $D_U$ of unlabeled instances
  Loop for k iterations to reach user satisfaction:
    • Train classifier $h_1$ from $D_L$.
    • Train classifier $h_2$ from $D_L$ using bagging.
    • Pick a random sample pool R from $D_U$.
    • Apply $h_1$ and $h_2$ to $R$ and compare their predictions.
    Let $\forall x \in R$ : $P_1(x)$ and $P_2(x)$ be the class probability
    Distributions estimated by $h_1$ and $h_2$ respectively and
    S be the set of instances from R such that:
        $\forall x \in S$ : $h_1(x) \neq h_2(x)$
                $Rank(x) = JS(P_1, P_2)$
    • Select $x = \arg\max_{x \in S} Rank(x)$, label it with the
      oracle and add it to the training set $D_L$.
```

Alg. 2 : Active learning to reduce bias and variance.

5. Empirical results

We applied our method to reduce zero-one loss in a series of experiments on used standard datasets from the UCI repository. We introduced uncertainty into the problem by assuming that at most 5% of the data is labeled, and used naïve Bayes classifier implemented by Weka (Witten & Frank, 1999) to do the experiments. First from the labeled data we trained two classifiers, one a naïve Bayes learner and the other by applying bagging with naïve Bayes as the base classifier to the set of labeled data and treating the result as the prediction of a single classifier. Then from the set of unlabeled data, we picked a random pool, applied algorithm 1 and recorded the expected bias, variance and loss over the test data before and after applying the algorithm. We stoped the experiment based on anytime strategy, which means the algorithm returns the best answer possible even if it is not allowed to run to completion and may improve on the answer if it is allowed to run longer. This strategy allows, in general, a trade off between solution quality and search time. In our case, the trade off will be between the amount of current loss and cost of using the oracle in labeling additional instance(s). The results were obtained on average from 3 runs of the algorithms on a dataset. We summarize the results of the experiments in Table 1 and Figures 1, 2, 3:

Table 1: Reduction of loss using MIS

Dataset	Initial Classifier	Active Learner
Iris	28.67	11.14
Auto.	39.94	37.83
Labor	34.17	10.9
Pima	31.66	26.34
Cpu.	15.75	15.58

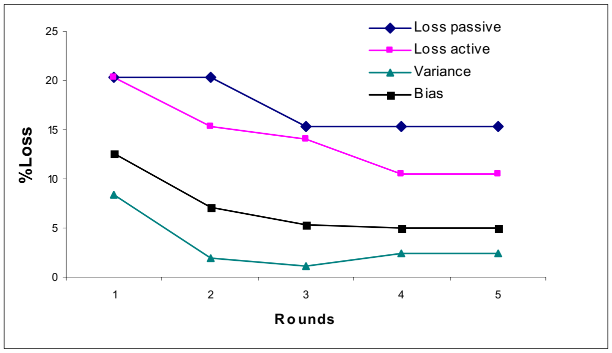

Figure 1: Reduction of loss, bias and variance of the learner in dataset Iris, using MIS.

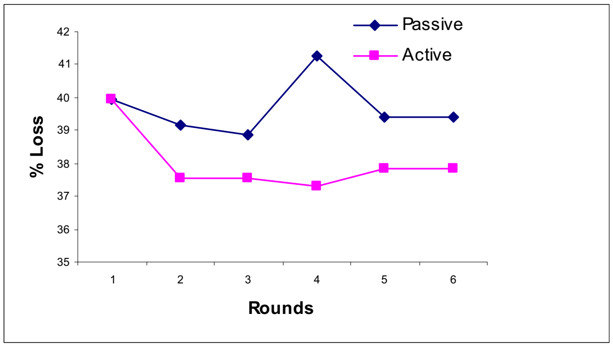

Figure 2: Reduction of loss of the learner in dataset Auto, using MIS.

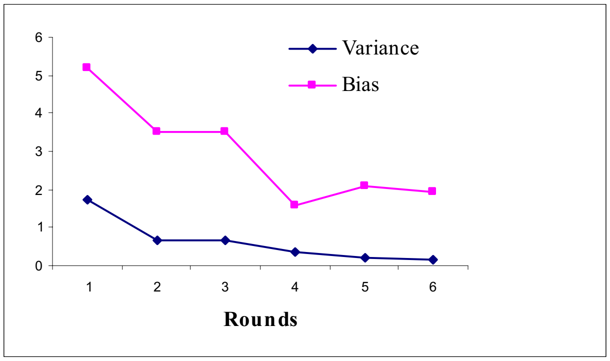

Figure 3: Reduction of bias and variance of the learner in dataset Auto, using MIS.

Next we tested algorithm 2 on data set Pima and compared the results of using different selection strategies, which we denoted by MIS, SIS, SIS-JSD, and passive learning as shown in Table 2 and Figure 4.

Table 2: Comparison of reduction of loss in dataset Pima using passive and active learning algorithms

Number Labeled	Passive Learning	Active Learning		
		MIS	SIS	SIS-JSD
23(initial)	31.66	31.66	31.66	31.66
30	30.95	28.69	27.95	27.15
38	31.75	27.35	26.34	26.47

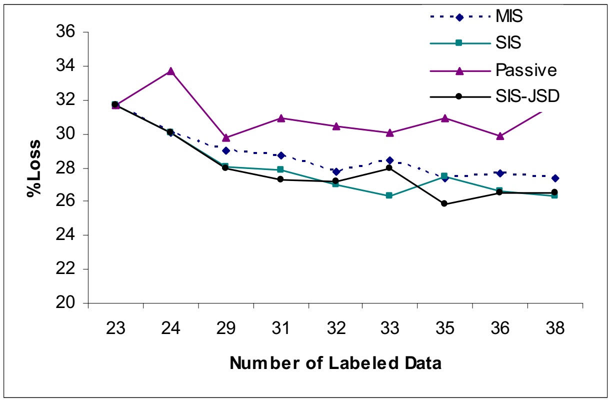

Figure 4: Reduction of loss for the dataset Pima.

Comparing the results we notice that both single instance selection policies SIS and SIS-JSD perform better than multiple instance selection policy, MIS. Our single instance selection policies will reduce loss more than MIS. Furthermore, we can reduce classification cost since we can end up labeling fewer instances to reach certain level of loss (SIS, SIS-JSD) than adding all the instance at once (MIS). Also, to compare the performance of SIS with SIS-JSD, we did a paired student t-test for means which resulted in the p-value of 0.8. This p-value is clearly larger than any reasonable significance level, therefore the difference between losses for SIS and SIS-JSD do not seem to be statistically significant.

6. Related work

The work of Melville and Mooney [22] focuses on using Jensen-Shannon divergence in active learning. They generate class probability estimates and compute two measures of utility for each example belonging to the unlabeled data: margins and JS-Divergence. They use one meta-learner with decision tree as their base learner.

Our work along with the work of Cohen focus on bias, variance reduction while other known approaches to active learning attempt to reduce other measures. The work of Mitchell and Tong focuses on reducing future error aim to reduce the version space [13] and [14]. Sometimes uncertainty is the criterion used as a metric to select examples [15]. None of these methods considers directly optimizing expected error on future test data.

In our algorithm 1, the selected unlabeled examples to label and include in the training set are those that are difficult to classify by both classifiers h_1 and h_2. In this sense algorithm 1 is similar to AdaBoost [24], but while boosting augments the training set with the machine-labeled ones in a weighted manner, we use an oracle to get the true value of the label.

In another method called co-training [1], authors propose to add to the PAC model a notion of compatibility between a concept and a data distribution. They postulate that if the target concept is compatible with the distribution given, this allows unlabeled data to reduce the class *C* of concept classes to the smaller set *C'* of functions in *C* that are most compatible with what is known about the distribution [1]. In this setting, we start with two weak classifiers, and through a number of iterations, reduce the high loss of each classifier resulting from labeling instances with low confidence, by getting help from the other classifier labeling those instances. Therefore, the two classifiers compensate for each other's weaknesses. For the problems which lack an explicit view factorization, Goldman and Zhou [8] and Steedman et al. [10] proposed two different learning algorithms to compensate for the multiple views required by the standard co-training. Ng and Cardie [4] proposed a single-view algorithm for bootstrapping co-reference classifiers, which is also applicable to the problems which lack an obvious feature split.

7. Conclusion and future work

We have proposed a method of learning from scarcely labeled data using active learning and shown empirically that our algorithms outperform the initial classifier or the approximation to Bayesian averaging model given the current training set. We applied our method to a number of data sets and proved that we can reduce the loss of the classifier through reducing its bias and variance via active learning. We also proposed two improvements to our method which can achieve certain level of loss with fewer queries to the oracle. Of course there can be other utility measures besides the ones that we chose to select a single instance or other similarity measures for probability distributions, we have chosen two, the rest can be the subject of future research.

In this paper we focused on reduction of 0-1 loss. In future we would like to continue working in two related areas of our particular interest first, given an estimated amount of loss after how many iterations can we stop? Second, given data sets with large amount of unlabeled data how can we select the best representative pool from the unlabeled data to further speed up the learner's convergence to a desired level of loss.

Previously, we have done some research in both areas [26], [27] and we plan to extend these studies further into semi-supervised learning using different types of loss functions and compare our proposed method with other algorithms in reducing bias and variance of the learner.

References

[1] A. Blum, & T. Mitchell,: Combining labeled and Unlabeled data with co-training. Proceedings of the Conference on Computational Learning Theory, 1998, pp. 92- 100.

[2] P. Domingos,: A Unified Bias Variance Decomposition. Proceedings of 17th International Conference on Machine Learning. Stanford CA Morgan Kaufmann, 2000, pp. 231-238.

[3] I. Muslea, S. Minton, C. A. Knoblock, : Active + Semi-Supervised Learning = Robust Multi-View Learning. Proceedings of the 19th International Conference on Machine Learning , 2002, pp. 435-442.

[4] V. Ng, and C. Cardie, Bootstrappng Conference Classifiers with Multiple Machine Learning Algorithms. Proceedings of the 2003 Conference on Empirical methods in Natural Language processing, Sapporo, Japan, pp. 113-120.

.[5] D. Cohn, Z. Ghahramani and M. I. Jordan,: Active Learning with Mixture Models. in R. Murray-Smith T. Johansen, eds. Multiple Model Approaches t Modeling and Control. Taylor and Francis. London , 1997.

[6] D. Cohen, Minimizing Statistical Bias with Queries.,1995 .

[7] S. Geman, D. Bienenstock, & R. Doursat: NeuralNetworks and the Bias/Variance Dilemma, NeuralComputation. 1992, Volume 4, pp. 1-58.

[8] S. Goldman, and Y. Zhou: Enhancing supervised Learning with unlabeled data. In Proceedings of International Conference on Machine Learning, 2000, pp. 327-334.

[9] M. Seeger,: Learning with Labeled and Unlabeled Data, Technical report, 2000.

[10] M. Steedman, M. Osborne, A. Sarkar, S. Clark, R. Hwa, J. Hockenmaaier, P. Ruhlen, S. Baker, and J. Crim: Bootstrapping statistical parsers from small datasets. In proceedings of the EACL. 2003.

[11] J. F. Gomez-Lopera, J. Martinez-Aroza,, A. M. Robles-Prez, & R. Roman-Roldan,: An Analysis of Edge Detection by Using the Jensen-Shannon Divergence. Journal of Mathematical Imaging and Vision, 2000, 13, pp. 35-56.

[12] T. Mitchell,: Machine Learning, McGraw Hill. 1997.

[13] T. M. Mitchell, : Generalization as Search, Artificial Intelligence, 18. 1982.

[14] S. Tong, & D. Koller,: Support Vector Machines Active Learning with Applications to Text Classification. Proceedings of the Seventeenth International Conference on Machine Learning. 2000.

[15] D. Lewis, & W. Gale, A Sequential algorithm for Training Text Classifiers.Proceedings of the International ACM-SIGIR Conference on Research and Development in Information Retrieval. 1994, pp. 3-12.

[16] N. Roy, and A. McCallum,: Toward Optimal Active Learning Through Sampling Estimation of Error Reduction. Proceedings of the 18th International Conference on Machine Learning. 2001.

[17] P. Domingos,: Bayesian Averaging of Classifiers and the Overfitting Problem. ICML, 2000.

[18] Y. Abramson, & Y. Freund,: Active Learning for Visual Object Recognition.

[19] K. Nigam, A. K. McCallum, S. Thurn, & T. Mitchell, Text Classification from Labeled and Unlabeled Documents Using EM. Machine Learning, 2000, pp. 1-32

[20] B. Shahshahani,& D. Landgrebe,: The Effect of Unlabeled samples in Reducing the Small Sample Size Problem and Mitigating the Hughes Phenomenon. IEEE Transactions on Geoscience and Remote Sensing, 1994, 32, pp. 1087-1095.

[21] T. Zhang, and F. Oles,: A probability Analysis on the Value of the Unlabeled Data for Classification Problems. International Joint Conference on Machine Learning. 2000, pp. 1191-1198.

[22] P. Melville, R. J. Mooney, :Diverse Ensembles for Active Learning, 21st International Conference on Machine Learning, In Proceedings of ICML-04, Banff, Canada, 2004, pp. 584-591.

[23] J. Lin,: Divergence Measures Based on the Shannon Entropy. IEEE Transactions on Information Theory, 1991, 37(1), pp. 145-151.

[24] Y. Freund, & R. E. Schapire, : A Decision Theoretic generalization of On-Line Learning and an Application to Boosting. Journal of Computer and System Science, 55(1), 1997, pp. 119-139.

[25] V. Castelli, and T. M. Cover,: The Relative Value of Labeled and Unlabeled Samples in Pattern Recognition with an Unknown Mixing Parameter. IEEE Transaction on Information Theory, 1996, 42(6), pp. 2102-2117

[26] I. Davidson, and M. Aminian,: Using Central Limit Theorem for Belief Network Learning. The 8th International Symposium on A.I. and Math, Fort Lauderdale, Florida, 2004.

[27] M. Aminian, : Co-training from an Incremental EM Perspective. The 5th International Conference on Intelligent Data Engineering and Automated Learning, Exeter, UK, 2004.