ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

Ensemble Prototype Vector Machines based on Semi Supervised Classifiers

S. Yamuna
M.Tech Student
Dept of CSE, Jawaharlal Nehru Technological University, Anantapur

Abstract: **-- In large number of real world applications the possibility of constituent volumes of unlabeled data is enormous. Moreover, the availability of labeled data is inadequate because of the expensive and annoying human interventions. The semi-supervised learning is a substitute of supervised learning model utilizes the small amount of labeled data for training the massive volumes of unlabeled collections is an adequate model FOE enhancing the learners' pursuance. In order to this Kai Zhang et al proposed a model that attempted to improve the Graph-Based Semi supervised Learning via Prototype Vector Machines. It uses scanty prototypes which are derived from data. Moreover, this mechanism will work effectively only on limited data samples. But, prediction of new data label from training data is more complex. The motivation gained from this model, an ensemble prototype vector machine for scaling classification performance that aimed to reduce the time and memory complexities of the kernel learning are used .The ensemble prototype vectors can handle large data sets without any complexity and for producing the new samples predictive analysis classification is performed on trained data. In predictive analysis, the decision trees are build on the training data for producing the new labels without any repeated factors. This ensemble model should achieve satisfactory classification performance.**

Keywords: **-- Semi-Supervised models, ensemble prototype vectors, Graph-based models, large data sets, predictive analysis.**

I. INTRODUCTION

The semi-supervised learning model uses labeled data during training to yield a precise classifier, has got sufficient volume of classifications from the community of research. It reduces expensive human efforts in supervised learning and faultiness in unsupervised learning models. Semi-supervised model uses labeled data(side-scoop). Generalized expectation criteria(GE), a model which is previously defined as an expectation regularization in McCallum and Mann(2007) is represented for using this data. Collection of semi-supervised learning models are represented by GE.where constraints are fit by reducing model branching perceived input data.

In semi-supervised learning graph based methods are needed for approximating the data as manifold, which represents data as weighted graph particles. Prototype vectors are used for classifying the graph based learning by clustering all samples in the data set. But in this model data samples are classified as clusters which yield different values beyond the limit of limited cluster volumes.

In this paper ensemble prototype vectors are applied to the labeled records which hardly reduce the overlapping, no independent sample sets and plots crystalline consistent estimates. Further, ensemble PVMS will assure precise scalability and little spoor at the time of testing. Prediction of new labels from unseen samples reduces the time complexity.

The rest of the paper is organized as follows. In section II some important semi-supervised learning models are reviewed. In section III application of ensemble prototype vectors and comparison with existing literature are represented. In section IV methodology of ensemble prototypes is defined, i.e., classification of data samples through prototype vectors using k-NN algorithm is defined. Section V reviews experimental results and section VI reflects some conclusions.

II. RELATED WORK

The Semi-supervised learning algorithms can be used in many cases without any problems chapelle et al(.,2006) proposed that semi-supervised methods does not overcome the supervised methods uniformly and their is no luminous winner in between these methods. This consequence emulates the theoretical support and the experimental index from a vast work period semi-supervised learning is disruptly un-predictable for expectation-maximization. In a simple consequence (Merialds, 1994) propone semi-supervised learning to enhance tagging of parts-of-speech in HMM and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

observes that scale down in efficiency due to expected-maximization with unlabeled samples. Ng and Cardie (2003) fails to apply the expected-maximization to improve the performance also cohen and cozmen (2006) employ use eases and finds EM work failures. Chen et al (2005) tried to merge manifold approaches with semi-supervised learning, since the techniques are very frangible in tuning specifications. Chawla and Blum (2001) graph based methods finds the frangibility in parameters tuning complications.

Instead of using semi-supervising learning algorithm with labeled and poor labeled scoop or side measurements, moreover none of them has labeling expressions like GE . Zhu et al, (2003) cities the graph based models uses the group of proportions testing to set the limits for label; propagation. Structural label distributions may also be used on unlabeled data for identifying the model structures (Schuurmans, 1997) , Burges and platt employ conditional harmonic merging for reducing the KL-divergence at each landmark among the currently insisted labels and the distribution employed by its neighbors. Wang et. al (2004) cite for a process for combining class segments with classification. By using the scheme of manifold and regularization several graph based semi-supervised learning algorithms are evolved like, self0training, co-training, SVMs, transductive SVMs.

These algorithms represent the graph based points in the form of the manifold assumption. Mare cleasen and the Frank De Smet (2014) employ the ensemble learning software for the SVMs which minimizes the training complexity by avoiding the repeated classes.

Hence ensemble prototype vectors are applied to the records while classification to improve the accuracy and with minimum time consumption.

III. ENSEMBLE PVMS

In existing system, the prototype vectors are used for graph based regularization in semi-supervised learning..Here the prototype vectors are the dispersed data types which are used for the replacement of the original data..Here the labels are classified by using mainly pair of prototypes:-

- low-rank approximation
- label-reconstruction prototypes

The low-rank approximation is used for reducing the graph-laplacian matrix without changing n*n kernel matrix. It chooses the set of ‘m’ samples from dataset which acts as landmark points. The label-reconstruction prototypes which evade the need of all labels from the whole dataset by using similarity between the samples. The k-means clustering is worn for sampling the landmark points which is linear in size. The prototypes are approximated by using square loss and hinge loss.

The k-means clustering is a subset of unsupervised learning algorithm that approximates the labels as clusters. If more number of prototype vectors are used in this process repeated samples will occur which leads to reduction of predictive accuracy.

To overcome this here data is classified as set of label records. The label records are approximated by using the ensemble prototype vectors which summarizes the unlabeled data. If noise occurs while processing the samples then k- means algorithm delineates the performance. To avoid this K-NN classifier algorithm is introduced.

IV. METHODOLOGY

In the existing literature, the labels are approximated as graph points using binary classifier that classifies the limited number label points which tends to increase in time consumption.

In proposed methodology K-nearest neighboring algorithm is used. The k-NN algorithm is a subset of supervised learning algorithm in which ensemble prototype vectors are classified by using nearest neighbor samples. The main computation of this algorithm is to classify the given data set into records. These label records are based on the size of data set, i.e., k and approximated by using these prototype vectors.

The prototype vectors are assigned to each data record for classifying them into label and unlabel samples. New label samples are generated by calculating similarity mean factors between nearest neighbors after grouping identical patterns.

While processing the data using prototype vectors noise may occur which leads to reduction in

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

prediction accuracy. In order to increase the performance i.e., improve the accuracy and reduce the noise pruning methodology is used.

K-NN ALGORITHM

1. Start the process by generating the set of label records.
2. Approximate the labels by using ensemble prototype vectors.
3. If unseen labels are generated then
4. Compare similarity factors between the samples and generate them as new samples.
5. End if
6. For generating the new class labels from unlabeled data calculate the nearest neighbouring distance and predict.
7. End.

The given data samples are classified as records by using the k-NN algorithm. The term k reflects the nearest neighbour values which is defined based on data set at training time. The nearest neighbour values are calculated by using "ecludiean distance" metric. i.e.,

$$nearest\ neighbour = \sqrt{\sum_{l=1}^{k} (p1 - p2)2}$$

P1and P2 are the nearest factors, after classifying the records using similarities between the samples and calculating the distance between them, the new labels are generated by using predictive analysis classification model.

The following figure represents the computation of K-NN algorithm.

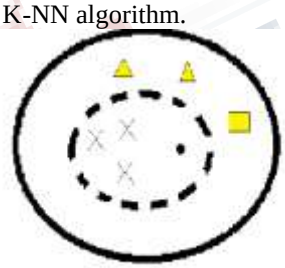

Fig1: K-NN algorithm computation

In the diagram the UN similar patterns are classified as new labels and transformed to predictive classification.

Predictive analysis:

For producing new labelled data sets in future based on previously classified labels predictive analysis is performed. In this process, unseen samples which observed at the time of classification are transformed to decision trees.

The decision trees will produce new label factors efficiently under classification by performing induction algorithm on data sets. This algorithm will improve the accuracy by avoiding the repeated samples with minimum complexity through pruning. While pruning the data, the labels are tested for reducing the noise in the data samples. The combination of K-NN algorithm with predictive classification yields good results.

V. SIMULATIONS AND ANALYSIS

A. The Dataset

The equitable of the process is to perform the semi supervised learning based classification of the data using ensemble prototype vector machines. To assess the scalability and clustering accuracy, we adopt the data set with labeled records.

B. Assessment metrics and strategy

The metrics that we considered to assess the accuracy of the classes formed by proposed model are precision, sensitivity, specificity and accuracy, which are estimated by using true-positives, false-positives, true negatives and false negatives. In order to obtain the true negatives and false negatives the adopted model is vector machines, which are often complexes towards process and resource utilization. Hence the time complexity and process complexity of the proposed algorithm also being assessed

Experimental Setup and Results

In performance analysis both assessment metrics resource and computational complicatedness also included in, a computer with i5 processor, 4GB ram and Nvidia 4GB graphics card (NVIDIA, 2015) used. The implementation was done in Java.

The input and obtained results were explored in table 1.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 3, Issue 10, October 2016

Total number of input records	Labeled : 1021, unlabeled :479
Total number of actual classes	14 from labeled records
Total number of initial classes	67 from all labeled and unlabeled records
Total number of predicted classes by proposed model	20 from all labeled and unlabeled records
True Positives	1007
False Positives	28
True Negatives	451
False Negatives	14
Precision	0.972947
Sensitivity	0.986288
Specificity	0.969892
Accuracy	0.972

Table 1: Input and observed metric values from the experiments

The performance of the model is assessed on a record set of size 1500. Among these records 1021 records already with known labels, which are notice to be fit into 14 classes. In order to assess the accuracy, the records of size 479 of divergent concepts, which are far different from the concepts of the labeled records are considered. The labeled records were considered as positives and unlabeled records were considered as negatives towards the actual classes defined. Further the classes predicted by Proposed Model were assessed, which is based on the association of the records given. The Metric values indicating that prediction of record associability under jaccard index (record relevancy to the cluster) by the Proposed Model is phenomenally significant (precision is 0.972947). The true positive Rate that indicates the true prediction of ratio of records for relevant cluster is also considerably high (sensitivity is 0.986288) for Proposed Model. The prediction rate of unrelated records to the defined classes is outstandingly high(specificity is 0.969892). The overall record clustering optimality by Proposed Model is find as good, moreover , the 97% of classes are clustered into the identical labels among the inclined input and empirical structure (accuracy is 0.972).

The complications of computing and source cost is also assessed, which is done under divergent count of initial classes as input. The time complexity observed to be linear for given initial classes as input (see fig 1). The memory usage of Incremental evolutionary genetic algorithm also being noticed as linear for given input classes (see fig 2).

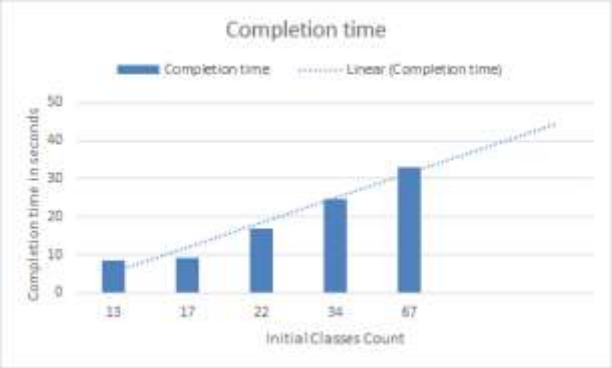

Figure 1: Ensemble prototypes completion time looks for factious volume of input classes

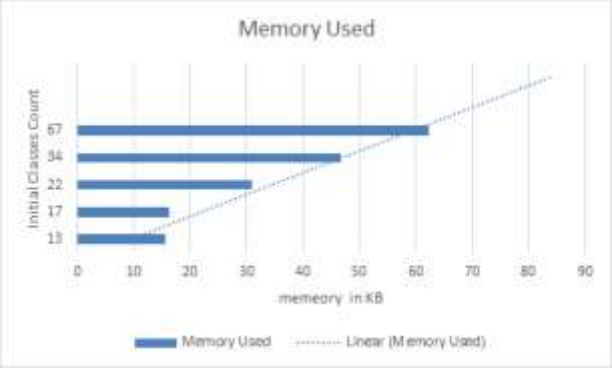

Figure 2: Memory worn for ensemble model

VI. CONCLUSIONS

In this paper, ensemble PVMs are proposed for scaling graph-based semi supervised learning models. prototypes are used for approximating the graph-based regulaizer, predictive model by predictive classification that decreases the time consumption, reduces the problem size and can employ model over large scale real world problems . speculations on numeral of actual input sets exhibits that the ensemble PVMS has appealing escalate behavior (linear size) and competitive performance. In further future work, various extensions of the ensemble PVM can be studied. For example, we can acknowledge alternative label reconstruction schema which acquire the local geometrical architecture into account. Moreover, the inter-relation between the labeled, unlabeled records, and the prototypes will be studied. New concepts and models are designed with rich possibilities for controlling the

hardware and memory representations.

REFERENCES

1) S. Abney. Understanding the yarowsky algorithm. Computational Linguistics, 30:3, 2004.

2) Y. Altun, D. McAllester, and M. Belkin. Maximum margin semi-supervised learning for structuredvariables. In NIPS, 2005.

3) R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. JMLR, 6, 2005.

4) S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: Taking random walks through the view graph. In WWW, 2008.

5) K. Bellare, G. Druck, and A. McCallum. Alternating projections for learning with expectation constraints. In UAI, 2009.

6) Y. Bengio, O. Dellalleau, and N. Le Roux. Label propagation and quadratic criterion. In O. Chapelle, B. Schlkopf, and A. Zien, editors, Semi-Supervised Learning. MIT Press, 2006.

7) L. Berger, V. J. Della Pietra, and S. A. Della Pietra. A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), 1996.

8) F. Bach and M. Jordan, "Predictive low-rank decomposition for kernel methods," in Proc. 22nd Int. Conf. Mach. Learn., Bonn, Germany, Aug. 2005, pp. 33–40.

9) S. Baluja, "Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data," in Proc. Adv. NIPS, vol. 11. 1999, pp. 854–860.

10) M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in Proc. Adv. NIPS, vol. 14. 2002, pp. 585–591.

11) M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," J. Mach. Learn. Res., vol. 7, no. 11, pp. 2399–2434, 2006.

12) S. Ben-David, T. Lu, and D. Pal, "Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning," in Proc. 21st Annu. Conf. Learn. Theory, 2008, pp. 33–44.

13) Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in Proc. 18th Int. Conf. Mach. Learn., San Francisco, CA, USA, 2001, pp. 19–26.