SURVEY ON SEMI SUPERVISED CLASSIFICATION METHODS AND FEATURE SELECTION

Neethu Innocent [1], Mathew Kurian [2]

[1] *PG student, Department of Computer Science and Technology, Karunya University, Tamil Nadu, India,* ***neethucii@gmail.com***

[2]*Assistant Professor, computer science and technology ,Karunya University, Tamil Nadu, India,* ***mathewk80@karunya.edu***

Abstract

Data mining also called knowledge discovery is a process of analyzing data from several perspective and summarize it into useful information. It has tremendous application in the area of classification like pattern recognition, discovering several disease type, analysis of medical image, recognizing speech, for identifying biometric, drug discovery etc. This is a survey based on several semisupervised classification method used by classifiers , in this both labeled and unlabeled data can be used for classification purpose.It is less expensive than other classification methods . Different techniques surveyed in this paper are low density separation approach, transductive SVM, semi-supervised based logistic discriminate procedure, self training nearest neighbour rule using cut edges, self training nearest neighbour rule using cut edges. Along with classification methods a review about various feature selection methods is also mentioned in this paper. Feature selection is performed to reduce the dimension of large dataset. After reducing attribute the data is given for classification hence the accuracy and performance of classification system can be improved .Several feature selection method include consistency based feature selection, fuzzy entropy measure feature selection with similarity classifier, Signal to noise ratio,Positive approximation. So each method has several benefits.

Index Terms: *Semisupervised classification, Transductive support vector machine, Feature selection, unlabeled samples*

----------***----------

1.INTRODUCTION

Data mining is a process of extracting information from large datasets and converts it into understandable format. It has major role in the area classification .Various data mining technique can be used for classification of several diseases It will helps for better diagnosis and treatment.

There are several classification methods to used by classifier. By using supervised classification method, only labeled data can be used but it is very expensive and difficult to obtain. This paper mainly concentrated on semi supervised classification method which use both labeled and unlabeled data[8]. Several semi supervised classification methods are transductive support vector machine, recursively partition model, cut edge and nearest neighbour rule, low density separation approach.

One major problem faced in the classification is due to large number of features of dataset. if there is thousands of features then it will affect the performance of classifier. This is one of the challenge in machine learning technique and is called feature selection[1],[7].By using this method only selected features will taken for classification so dimension of data will reduced, computational cost can be reduced, performance of classification can be increased . Several feature selection methods included in this paper are consistency based feature selection method, consistency based feature selection, signal to noise ratio , positive approximation, Fuzzy entropy measure feature selection with similarity classifier, positive approximation based on rough set theory.

2.SEMISUPERVISED CLASSIFICATION

In semisupervised classification method both labeled and unlabeled samples is used . Unlabeled samples will obtained very easily but labeled samples are very expensive.

2.1.Transductive SVM

TSVM it is an iterative algorithm which include unlabeled samples in training phase. Ujwal malik et.al proposed a transductive procedure in which transductive sample is selected through a filtering process of unlabeled data and an algorithm is proposed[1].

This algorithm use input as both labeled and unlabeled samples. Algorithm starts with training the SVM classifier which is having a working set T(0).Working set will be equal to labeled set . The unlabeled data which fall into the margin will have more information, in which some data fall into negative side and is called negative transductive sample and unlabeled data fall into positive side is called positive

transductive samples. Samples with accurate labeling ,informative samples and samples which are near to margin will be selected and some will be residing in upper side and in lower side will be assigned as +1and -1 respectively. Selected transductive sample is added to the training set. So by using this method the accuracy of classification can be increased and cost also reduced.

2.2.Semi-supervised based logistic discriminant procedure

Shuichi kawano et.al proposed a non linear semi-supervised logistic discriminant procedure which is based on Gaussian basis expansions with regularization based on graph [9] .Graph laplacian is used in regularization term is one of the major technique in graph based regularization method .It is based on degree matrix and weighted adjacency matrix . Weighted matrix M will be a n x n matrix. To select values of several tuning parameters they derive a model for the selection criteria from Bayesian and information theoretic approach . Laplacian graph are applied easily to analyze high dimensional or complex dataset in both labeled and unlabeled data set. This method also reduce error rate of prediction.

2.3.Self training nearest neighbour rule using cut edges

One of the common technique used in semisupervised method is self training In this case first classifier will train with labeled samples and then it will include unlabeled samples and added with training set. Classifier teaches s itself by using its prediction. But this method can cause several problems like misclassification, much noise in labeled data and also cause error reinforcement. Y.Wang et.al proposed a method to solve this problem and is called self training nearest neighbour rule using cut edges[2] .This method is to pool both testing samples and training sample in an iterative way. There are two aspects .

- Firstly maximum number of testing samples must classify to positive or negative. The extreme output must be of lower risk. If maximum number of class is obtained then construct a label modification mechanism which utilizes cut edges in the relative neighborhood graph.

- Secondly to employ cut edge weight for semisupervised classification technique. So by using cut edge it reduces classification error, error reinforcement and also improves the performance.

2.4.Low density separation approach

The algorithm of low density separation use cluster assumption which use both labeled and unlabeled data is used. Based on cluster assumption the decision boundary must lie in low density region and should not overlap high density region[1][4]. There are two procedure to keep the decision boundary in the low density regions between clusters. First, it obtain the graph based distance that give importance to low density region. Secondly it avoid high density region and obtain decision boundary by optimizing transductive SVM objective function. Hence these two procedures will combine. So LDS can achieve more accuracy compared than traditional semisupervised method and SVM.

Table 1: Comparison table for semisupervised classifications

Transductive svm	Accuracy of classification increased
semi-supervised based logistic discriminant procedure	Reduce error rate
Self trainng nearest neighbour rule using cut edges	Improve performance
Low density separation approach	Accuracy of classifier increased

3.FEATURE SELECTION

Most of the dataset will be of huge size. So feature selection is used to reduce the dimensionality of large data[10]. After reducing features the sample data is given to classifier . For example Gene expression data set is a large data set with thousands of gene data. So feature selection method is used to reduce the size of data and it can increase the performance of classifier. Several feature selection method are explained in this section.

3.1.Consistency based feature selection

In this method only relevant features are selected and also inconsistency measure is calculated [1],[5]. For example a pattern is having more than one matched instances but they are residing in different classes considered as it is inconsistent .So inconsistent features are removed. The forward greedy algorithm is used in this method the step of this algorithm is shown in the fig 1 [6]. Several steps are

- Features of next candidate subset is generated based on generation procedure
- Candidate function is evaluated based on evaluation function
- Decide a stopping criteria when to stop

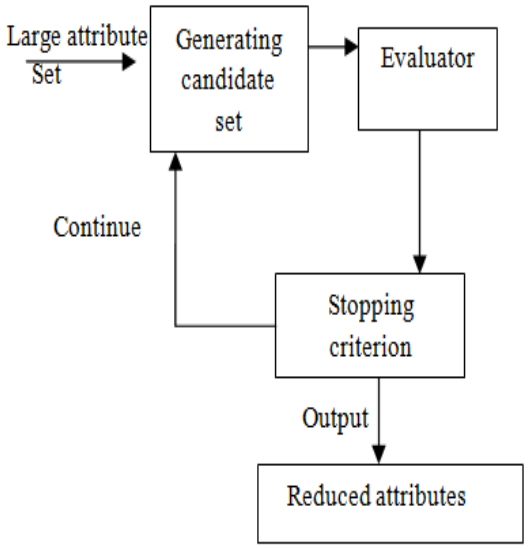

Fig 1.Forward Greedy Algorithm for attribute reduction

3.2.Fuzzy entropy measure feature selection with similarity classifier

In this method features is selected by fuzzy entropy measure and also with the help of similarity classifier[11]. The main principle behind similarity classifier is to create an ideal vector Vj=(-Vj(S1),……,Vj(Fn)) which represent the class j. This ideal vector is calculated from a the sample Yj of vector Y=(Y(S1),….Y(Sn)) which belongs to class Cj. After the calculation of ideal vector the similarity of vector V and sample Y is calculated, it is represented as S(Y,V). Value of similarity will be 1 in the ideal case if sample belong to class j. From this point view the entropy value is calculated to select features. If similarity value is high then entropy value will be very less and vice versa. Fuzzy entropy value of several features is calculated using by using sample vector and ideal vector which is used for classification. Features with highest entropy value are removed and lowest entropy value is selected. Classification accuracy is increased and computational time is reduced.

3.3.Signal to noise ratio

Signal to noise ratio is calculated by using the equation[1] , [3]

$$SNR= \frac{(m1-m2)}{(\sigma1-\sigma2)} \qquad (1)$$

m1and m2 are mean and σ1 and σ 2 are standard deviations. For example take the gene expression data then calculate the SNR value of based on gene expression level. Then arrange in descending order and select top ten features. This method will enhance the accuracy of classification.

3.4.Positive approximation

Existing heuristic attribute reduction has several limitations . Yuhua Qian proposed a method called positive approximation based on rough set theory[6]. The main objective of this method is to select some property of original data without any redundancy. There will be more than one reduct. But only one reduced attribute is needed so a heuristic algorithm is proposed based on significance measure of the attribute. This method is somewhat similar to greedy search algorithm. But some modification is proposed here significance measure is calculated . Until the reduct set is obtained the attribute with maximum significance value is added at each stage. The result of positive approximation of attribute reduction shows that it is an effective accelerator. There are three speed up factor in positive approximation based feature selection :

- One attribute can select more than one ineach loop. So this will helps to provide a restriction in the result of the reduction algorithm.
- Reduced computational time due to attribute significance measure.
- Another important factor in this algorithm is size of the data is reduced and time taken for the computation of stopping criteria is also reduced to minimum.

4.CONCLUSION

This paper based on review of various semisupervised classifications. Each classification method has its own advantages and disadvantages Low density separation approach used for classification can overcome the problems of traditional SVM. The other method transductive SVM can increase accuracy of classification. Logistic discriminant procedure and self training nearest neighbour rule using cut edges can reduce the error rate and misclassification. Survey on several feature selection methods is analyzed in this paper. Consistency based feature selection method reduce the inconsistent feature and increase the the performance. The another method fuzzy entropy measure feature selection with similarity classifier can increase the accuracy and reduce computational time. In signal noise ration feature is selected based on SNR values of each attribute and arranged in descending order then top ten features are selected . By using positive approximation reduction method the selected attribute will not have redundant values.

REFERENCES

[1]Ujjwal MauliK, Anirban Mukhopadhyay and Debasis Chakraborty “gene-expression-based cancer subtypes prediction through feature selection and transductive SVM” IEEE transactions on biomedical engineering, vol. 60, no. 4, april 2013.

[2]Yu Wang , Xiaoyan Xu, Haifeng Zhao , Zhongsheng Hua "semi-supervised learning based on nearest neighbor rule and cut edges" Knowledge-Based Systems 23 (2010) 547–554.

[3]Debahuti Mishra, Barnali Sahu "Feature selection for cancer classification: a signal-to-noise ratio approach" International Journal of Scientific & Engineering Research, Volume 2, Issue 4, April-2011 D. C. Koestler, C. J. Marsit, B. C. Christensen, M. R. Karagas, R. Bueno.

[4]O. Chapelle and A. Zien, "Semi-supervised classification by low-density separation," in Proc. 10th Int. Works. Artif. Intell. Stat., 2005, pp. 57–64

[5]M. Dash andH. Liu, "Consistency based search in feature selection," Artif.Intell., vol. 151, pp. 155–176, 2003.

[6]Y. Qian, J. Liang, W. Pedrycz, and C. Dang, "Positive approximation:An accelerator for attribute reduction in rough set theory," Artif. Intell.,vol. 174, pp. 597–618, 2010.

[7]S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," oinformatics, vol. 23, no. 21, pp. 2859–2865, 2007.

[8] O. Chapelle, V. Sindhwani, and S. S. Keerthi, "Optimization techniques for semi-supervised support vectors," J.Mach. Learn. Res., vol. 9, pp. 203–233, 2008.

[9] Shuichi Kawano · Toshihiro Misumi · Sadanori Konishi "Semi-Supervised Logistic Discrimination Via Graph-Based Regularization" Neural Process Lett (2012) 36:203–216

[10] A. Blum and P. Langley, "Selection of relevant features and examples in machine learing," Artif. Intell., vol. 97, no. 1/2, pp. 245–271, 1997

[11] Pasi Luukka "Feature selection using fuzzy entropy measures with similarity classifier" Expert Systems with Applications 38 (2011) 4600–4607

BIOGRAPHIES

Neethu Innocent pursuing her M.Tech in Software Engineering from Karunya University, Tamilnadu, India. She received her Bachelor's degree from MG university in Computer Science and Engineering from Kerala.

Mathew Kurian,has finished his M.E in computer science and engineering from Jadavpur University,Kolkatta and currently he is working as Assistant Professor in Department of Computer Science and Engineering in Karunya University.Previously,he worked as Software Engineer with Aricent Technologies.He is currently doing his PhD Degree in Data Mining.