Small Data Set Classification By Extending Attributes Information For Improving Classification Accuracy

Miss. Rina S. Patil[1], Mr. Kishor P. Jadhav[2], Mr. Harshal .S.Sangle[3]
[1, 2, 3] Dept of Information Technology
[1, 2, 3] Sanjivani KBPP Kopargaon

Abstract- *TWhen we are using small data set then main issue is data quantity, because robust classification performance will not be calculated due to this inadequate or small data. So extracting more efficient information is the new research area which is developing now a days. Considering this new field this paper proposes a new attribute construction method. This method converts original attributes in higher dimensional feature space. This will help to extract more attribute information using classification oriented fuzzy membership function which is available in similarity-based algorithm. To examine performance of the proposed method seven data sets having different attribute sizes are used. From result it is seen that proposed method has efficient classification performance than principal component analysis, kernel independent component analysis, and kernel principal component analysis.*

Keywords- Data mining, classification, small sample set, mega trend diffusion, attribute construction, support vector machine.

I. INTRODUCTION

In this competitive market, there are many situations when organizations must work with small data sets. For example, with the pilot production of a new product in the early stages of a system, dealing with a small number of VIP customers, and some special cancers, such as bladder cancer for which there are only a few medical records available. Data quantity is the main issue in the small data set problem, because usually insufficient data will not lead to a robust classification performance. How to extract more effective information from a small data set is the new research area now [1]. According to the computational learning theory, sample size in machine learning problems has a major effect on the learning performance. Faced with this issue, adding some artificial data to the system in order to accelerate acquiring learning stability and to increase learning accuracy is one effective approach. In virtual data generation, the prior knowledge obtained from the small training set given helps to create virtual examples to improve in pattern recognition. Analysts hope to acquire more training data before conducting a learning task, since learning based on a small data set faces the problem of insufficient information.

Shawe-Taylor et al. [2] proposed Probably Approximately Correct (PAC) to determine the minimum sample size required for the necessary accuracy. Muto and Hamamoto [3] stated a rule for the size of sample data based on the ratio of the training sample size to the number of attributes. Many researchers proposed various linear models for analyzing small data sets. Schwarz [4] derived the Schwarz Information Criterion (SIC) using a Bayesian perspective for model selection, where the Bayes solution is used to choose the model with the largest posterior probability of being correct. In machine learning problems, small sample size plays important role, because without these few samples information will not be complete. For example, with a classifier, it is hard to make accurate forecasts because small data sets not only make the modeling procedure prone to over fitting, but also cause problems in predicting specific correlations between the inputs and outputs. Virtual sample generation approach was proposed for enhancing classification performance for small data set analysis, but original idea was proposed by Niyogi et al. [5]. Support vector machine (SVM), are commonly used classifiers..

Now a days the manufacturing environment changes promptly owing to globalization and innovation. It is noteworthy that the life cycle of products consequently becomes shorter and shorter. Although data mining techniques are widely employed by researchers to extract proper management information from the data, scarce data can only be obtained in the early stages of a manufacturing system. From the view of machine learning, the size of training data significantly influences the learning accuracies. Learning based on limited experience will be a tough task. Consequently, investigators always want to acquire more training data to implement learning tasks; nonetheless for small-data-set learning, the problems encountered seriously results from insufficient information.

When learning with small data sets, fuzzy theory is another way to overcome the insufficient information whose membership function provides various degree of data ambiguity. Li et al developed the data fuzzification technology based on fuzzy theory for improving the scheduling knowledge of FMSs, which only contain a small data set for

learning. Therefore, in order to fully fill the information gaps, a technique called mega diffusion was substituted a sample set for diffusing samples one for one. Furthermore, a data trend estimation concept is combined with the mega diffusion technique to avoid over-estimating. This technique, which combines mega diffusion and data trend estimation, was called mega-trend-diffusion.

II. IMPROVING CLASSIFICATION PERFORMANCE

Three general attribute space transformation approaches are: attribute selection, feature extraction, and attribute construction. Attribute selection is the process of choosing the subsets of attributes for learning [6]; feature extraction is the process of turning general representations into more specific ones [7]; and attribute construction is creating effective new attributes for knowledge modeling [8].

2.1. Selecting attribute

Variables whose variance is less than measurement noise are not important to the model. Conventional methods of feature selection involve evaluating different feature subsets using some indexes and selecting the best among them [1]. The index usually measures the representation capability in the classification or clustering analyses, depending on whether the selection process is supervised or unsupervised [9].

2.2. Extraction of the Feature

This technique has ability to project the original features into a lower feature space to reduce the number of data dimensions and improve analytical efficiency. This techniques are classified in two types: linear and nonlinear. Linear methods, like principal component analysis, reduce dimensionality by performing linear transformations on the input data. It also discover globally defined flat subspace. These methods are most effective if the input patterns are distributed more or less throughout the subspace [1].

2.3. Construction of Feature

It is process of creating a new description using existing description of an object. Generally, feature construction is the creation of new features which are currently described implicitly by other attributes. The difference between attribute construction and feature extraction is that the latter will usually result in significantly fewer features being presented in the data set [10], while the former adds features to it.

The method used for pilot run modelling of manufacturing systems is called Bootstrap Method where initial data set is small. Using the limited data obtained from pilot runs to shorten the lead time to predict future production is considered in this method. Although, artificial neural networks are widely utilized to extract management knowledge from acquired data, sufficient training data is the fundamental assumption. Unfortunately, this is often not achievable for pilot runs because there are few data obtained during trial stages and theoretically this means that the knowledge obtained is fragile. The bootstrap implies re-sampling a given data set with replacement and is used for measuring the accuracy of statistical estimates. The bootstrap is applied to generate virtual samples in order to fulfill the data gaps. But, the bootstrap procedure is executed once for each input factor not to resample a job. With the help of this method the error rate can be significantly decreased if applied to a very small data set [9].

The method which extend the attribute information on following way as, collecting the data , building MTD functions, and computing the overlap area of the MTD functions, and then moves on to class-possibility attribute transformation, attribute construction, attribute merging, and finally SVM model building is known as Mega Trend Diffusion Function (MTD). After data collection, this method begins to build the triangular fuzzy membership function (called the megatrend diffusion function) for each class in every attribute, and then computes the overlap area of the fuzzy membership functions of each class.

When the overlap area of membership functions is high, this means the class-possibility method cannot judge the classes very clearly, and the attribute will thus be analyzed by attribute construction. The attributes with low overlap area of membership functions will be examined by the class-possibility method, in which the class-possibility values are computed using fuzzy membership function called mega trend diffusion function. After constructing the attributes by class possibility and synthetic attributes, both the tables are merged to form a data set with high dimensions which can be applied to the classier to build classification model [3].

Pros and Cons of previous system

1. Data quantity is the main issue of the small data set because it create problem in classification performance.
2. Extracting effective information from small data set was also a problem of concern.
3. With small data set it is not possible to accurately forecast because small data set make modelling

procedure difficult and also cause problem in specific correlation between input and output.

4. Information in data of small size is scarced and has some learning limit.
5. Decision is hard to make under the limit data condition.
6. The computational learning theory develops mathematical models to describe the learning data size and number of training in machine learning.
7. It can also be applied to small data set learning. However, it still leaves some practical problems.
8. Although it offers a probably approximately correct (PAC) model to estimate the relation about predict accuracy and sample size, it is hard to calculate the sample space in the model. However, it indeed builds a theoretical model to describe the machine learning problem.

III. IMPLEMENTATION DETAILS

3.1 Building a Mega-Trend Diffusion Function for each Class

The MTD function was proposed by Li et al. [11] to deal with the small data set problem for scheduling strategies in early flexible manufacturing systems, and it is a triangular fuzzy membership function. The main purpose of the MTD function is to generate virtual samples to solve the problem of insufficient data in small data set analysis. Li et al. used the membership function in fuzzy set theory to calculate the possibility values of virtual samples instead of the probability in statistics to avoid the normal distribution assumption.

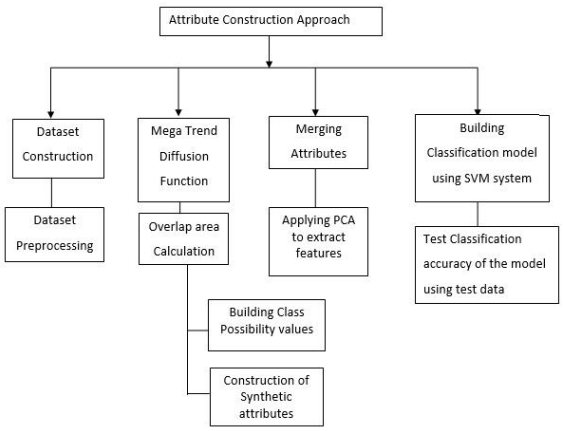


Figure 1: Module Breakdown structure

Fig. 2 shows the concept of the fuzzy theorem applied to the MTD function. The triangle is the membership function, and the height of samples m, and n are the possibility values of the membership function, denoted as M(m) and M(n).

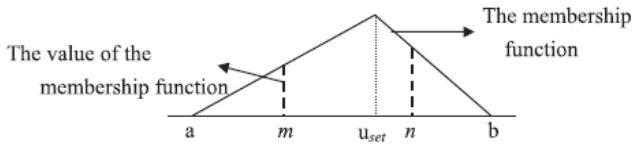


Fig.2. MTD Function

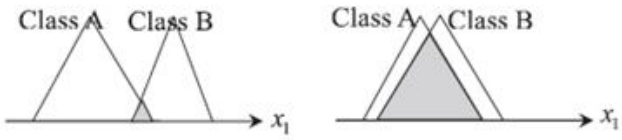


Fig.3.(a) Low Overlap (b) High Overlap

3.2 Computing the Overlap Area of MTD function T

After building the MTD function for each class in every attribute, finding the overlap area of MTD functions is an important step for data information extension. Fig. 3 shows an example of MTD function overlapping. Fig 3 (a) and 3 (b) show the low and high overlap of MTD functions for two classes, A and B, in attribute x1, In attribute x1, the area of overlap of the two classes is low, meaning that attribute x1 is an informative classification index because any point in attribute x1 can easily be classified into the correct class. Similarly, when the overlap area is high, the ability to place any point into the correct class will decrease. Hence, for attributes for which the area overlap is low, this study will add the class-possibility values as new attributes to the data set to extend the data dimension into a higher feature space to enhance the classification accuracy. For the attributes with a high overlap area, the attribute construction method will be introduced to construct new attributes substituting the original ones

3.3 Building Up the Fuzzy based Transformation Function

For every sample xi € X with M attributes, use the transformation function to extend the attribute from M attributes into M x (K+1) dimensions.

Separate the training data into two sets: classes A and B.

Apply the MTD technique to compute the membership grade of each attribute in each class.

Use the fuzzy-based transformation functions to extend xi into a high dimension.

Based on the MTD distribution, M(x), considering the classification problem with k-class, the transformed x produced by the fuzzy based transformation is:

$$\Psi(x) = (x, M1(x), M2(x),\ldots\ldots, M2(x)), x \in R$$

Fig. 4 shows a two-class problem. The triangle on the left side with a solid line represents the transformation function for class 1, denoted as M1(x). The triangle with a dotted line represents the transformation function for class 2, denoted as M2(x).Thus, for the two-class one-attribute classification problem, the transformed data are Ψ(x)= (x, M1(x), M2(x)).

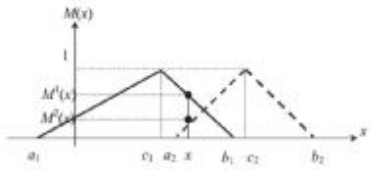


Fig.4.Building up the Fuzzy-Based Transformation Function

Assuming that we have a sample set X= (x1,t1), (x2,t2),….,(xN, tN), with K classes where each sample xi, i=1,2….,N , in X has M attributes (means xi =(x1,…, xM)), and t, is the target value of xi.

Step 1: Separate the sample set X into K subsets by its corresponding class target denoted as X {X˜1, X˜2, …, X˜k}, where X˜k (x1, k), (x2, k),. . . . , k=1,….,K, $0 < S \leq N$.

Step 2: Starting with class 1, the value of attribute 1, x1, X˜1 of the samples in X˜1 is denoted as x1i1, i=1,2,…S. Use this value to derive the transformation function for attribute 1, M1(x1), and repeat this computation for every attribute to obtain M1(xj), j=1,2,...,M, as shown in Fig.5 Iterate this step K times for each class to build up K × M transformation functions, Mk (xj), j=1,2, . . .,M , k=1,….,K.

Step 3: Starting with attribute 1, x1, for all samples in X. The transformation function of x1 is set as (x1, M1(x1), M2(x1),…., Mk(x)). Repeat this step M times to get (xj, M1(xj), M2(xj),….,Mk(xj)), where j=1,2,…,M. Hence, for every sample xi Є X with M attributes, use the transformation function to extend the attribute from M attributes into M × (K +1) dimensions. In this step, the computational complexity of the fuzzy-based transformation that transform the original data set into the new space is O(NxKM).

Step 4: After all samples have been transformed ,use PCA to extract the features. In this step, the computational complexity of estimating the PCA is

O((M X (K +1))2N) + O((M X (K + 1))3)

Step 5: Input the data set with the features extracted by PCA into the SVM learning model. In this step, SVM is used as the classifier. In Steinwart's research, the computation the complexity of SVM is O(N nSV + nSV3), where nSV denotes number of support vectors for a problem.

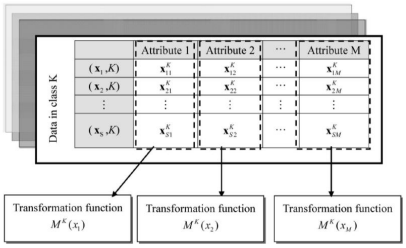


Fig.5.Transformation function building process for each attribute of samples in class 1

3.4 Attribute Construction

This explains attribute construction process for which the class overlap area is high. First, considering that two high overlap attributes may or may not have high correlation, the Pearson correlation coefficient is employed to further confirm the similarity between any pair of attributes. This study will then construct new attributes, named synthetic attributes, using the attributes that have a high correlation.

3.4.1 Compute the Correlation Matrix

In statistical analysis, the correlation coefficient plays an important role in measuring the strength of the linear relationship between two variables. In the field of computation, the correlation coefficient is one of the most well-known criteria for measuring similarity between two random variables. The correlation coefficient plays an important role in measuring the strength of the linear relationship between two variables

3.4.2 Attribute Combination with Highly Correlated Attributes

After computing the correlation of each pair of attributes, in this section will combine those with a high correlation value by using three constructive operators. Gomez and Morales proposed seven constructive operators: A+B, A*B, A-B, B-A, A/B, B/A, and A2. This study extracts only three of nonlinear operations, A*B, A/B, and B/A, as the constructive operators for the chosen attributes. The rest of the linear operators are substituted by PCA, because the main purpose of PCA is to extract the features by maximizing the variance of the linear combination for all attributes. Hence,

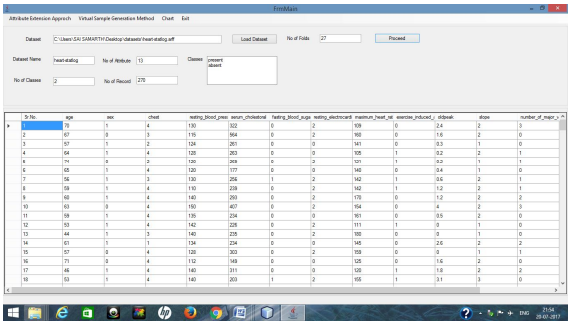
three operators, A+B, A-B, and B-A are considered in this study as redundant to PCA. In addition, the operator A2 is also not considered here.

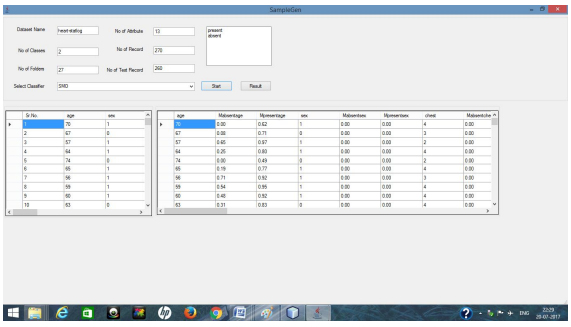
3.5 Build SVM Model

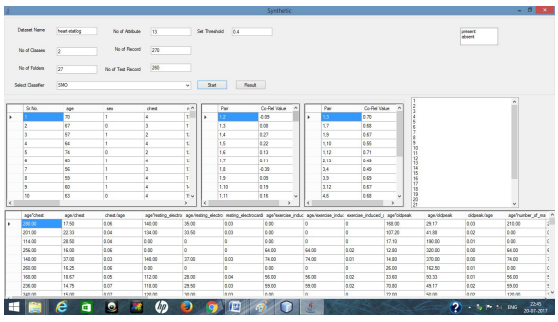
SVM classifier with a Gaussian kernel is used for building a classification model after preprocessing data set by the attribute construction method.

3.6 Snapshots of the Proposed Model

In this window, user is provided with the option of loading a training data set. User can select the data pre-processing method from the menu bar. After loading the data set and selecting the method, a classifier is chosen. After entering the size of each fold, the data set is divided into number of training and test set and training set is used for building the classifier. The class label of records is predicted for test set. Average accuracy is computed and displayed to user.







IV. RESULTS

4.1 Dataset

Table 1 shows information of each data set in terms of number of records, number of attributes, and number of classes.

Table 1. Data set Information

No.	Name of Dataset	No. of Attributes	No. of Instances	No. of Classes
1	Australian	14	690	2
2	Bladder	8	18	2
3	BUPA	6	345	2
4	Glass	10	214	7
5	Heart	13	270	2
6	Iris	4	150	3
7	Pima	8	768	2
8	Wine	13	178	3

Table 2. Attribute extended for each dataset

Name of Dataset	No of class possibility value	No of synthetic attributes
Australian	28	12
Bladder Cancer	16	21
Heart stat-log	26	6
Liver disorder	12	18
Glass	66	21
Iris	16	12
Pima	16	6
wine	50	18

V. CONCLUSION

A small training dataset usually leads to low learning accuracy with regard to classification of machine learning, and the knowledge derived is often fragile, and this is called small sample problem This paper aimed at obtaining a high classification accuracy by adding more information to small data set. For this purpose, the different attribute extension approaches are investigated. It is observe that accuracy of the system increases by increasing the attribute information, after performing several experimentations. By generating class possibility values for multi class problem, the accuracy is improved significantly. The system outperformed the existing data preprocessing methods for multi class problem.

REFERENCES

[1] Der-chiang Li and Chiao Wen Liu, "Extending Attribute Information for Small Data Set Classification," IEEE Transaction On Knowledge And Data Engineering.Vol.24,No.3,March 2012.

[2] D.C. Li, C.S.Wu, T.I Tsai, and Y.S. Lina, "Using Mega-Trend-Diffusion and Artificial Samples in Small Data Set Learning for Early Flexible Manufacturing System Scheduling Knowledge," Computers and Operations Research, vol. 34, pp. 966-982, 2007.

[3] M.A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 15, no. 6, pp. 1437-1447, Nov./Dec. 2002.

[4] [48] R. Thawonmas and S. Abe, "A Novel Approach to Feature Selection Based on Analysis of Class Regions," IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 27, no. 2, pp. 196-207, Apr. 1997.

[5] C.F. Huang and C. Moraga, "A Diffusion-Neural-Network for Learning from Small Samples," Int'l J. Approximate Reasoning, vol. 35, pp. 137-161, 2004.

[6] Chongfu Huang and Claudio Moraga, "The Generalized-Trend-Diffusion modelling algorithm for small data sets in the early stages of manufacturing systems", European Journal of Operational Research, 207, (2010),121-130.

[7] Tung-I Tsai, Der-Chiang Li,"Utilize bootstrap in small data set learning for pilot run modelling of manufacturing systems, "Expert Systems with Applications, 35, (2008), 1293-1300.

[8] C.L. Blake and C.J. Merz, "UCI Repository of Machine Learning Databases," Dept. of Information and Computer Science, California, Irvine, 1998.

[9] H. Liu and H. Motoda, Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer Academic Publishers, 1998.