International Journal for Research in Engineering Application & Management (IJREAM)
ISSN : 2454-9150 Vol-07, Issue-05, AUG 2021

Small sample dataset classification by extending attributes

[*]Yogesh A. Pawar, [#]Prof. Manish Rai

[*]M. Tech. C.S.E, [#]Professor, Department of Computer Science and Engineering, RKDF C.O.E. BHOPAL (M.P.), India. [*]pawaryogesha@gmail.com, [#]manishrai2587@gmail.com

Abstract- **This paper describes small data set classification by extending attribute information. Classification is one of the important techniques in data mining. It is very difficult to build a classification model when the data size is small. As well as a decision is hard to make under the limit data condition and data characteristics such as data distribution, mean, and variance are unknown.**

In this paper, an attribute construction is proposed that how extracts more effective information from a small sample set is thus of considerable interest. The proposed method starts from collecting the data, building MTD function, 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.

Index Terms*—Data mining, classification, small sample set, mega trend diffusion, attribute construction, support vector machine.*

I. INTRODUCTION

Classification systems play an important role in business decision-making tasks by classifying the available information based on some criteria. As classification systems become an integral part of organizational decision support systems, adaptability to variations in data characteristics and dynamics of business scenarios becomes increasingly important. In recent years, there has been a tremendous growth in the studies of the small data set learning methods in the condition of the paucity of data. Without double, information in data of small size is scared and have some learning limit. As well as a decision is hard to make under the limit data condition and data characteristics such as data distribution, mean, and variance are unknown [2].

To rapidly and continually improve the standard of product and services, companies must meet increasing demand of customer in highly competitive global market. In real world organization have to face many situations when they have limited data that is in other words we can say they have small data set. From the computational learning viewpoint, small sample sizes are very important in machine learning problems, because too few samples will contain incomplete information. For instance, with a classifier, it is hard to make accurate forecasts because small data sets not only make the modelling procedure prone to over fitting, but also cause problems in predicting specific correlations between the inputs and outputs. If we have small or limited data set then it may contain incomplete information. Small sample sizes are very important in machine learning problems from computational view point. This proposed work analyzes the characteristics of small data set learning. The Mega-trend diffusion method for small data set learning is applied mainly. Suppose we take an example of any disease, in this case we have limited number of patients that is we have only few medical records. So for our sample production (data set) we will have this limited record set.

The proposed work introduces an attribute construction approach that can be applied to small dataset where class labels are available. After collecting data, the proposed work begins to build the triangular fuzzy membership function, MTD (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 [3].

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 by the membership functions built for every attribute, and are used as new attributes added to the original attribute set. In the attribute construction method, the attributes with a high overlap area of membership functions are collected, and the correlation coefficient matrix is computed. Each pair of attributes with a high correlation coefficient will be used to

 DOI : 10.35291/2454-9150.2021.0437

construct the new attributes, viz. synthetic attributes, and principal component analysis is used to extract the useful attributes. Finally, the original attributes are merged with the attributes built up by class possibility, and the attributes created by attribute construction methods as the learning set for classier to form the classification model.

Once the model has been formed the test data set can be applied to the model which then predicts the class label of the records present in the test data set. The key advantage of the proposed method lies in the preservation of the original data structure and the addition of discovered classification information to the analysis, thus improving the learning accuracy for small data set modelling.

The rest of the paper is organized as follows: Section 2 describes approaches to improving classification performance, along with pros and cons. Section 3 describes system architecture and details of data used. Section 4 includes the sample set and results and finally, we draw conclusion and acknowledgement in Section 5.

II. LITERATURE SURVEY

Many studies have been conducted to improve the model accuracy of small data sets analysis. The main reason why small datasets cannot provide enough information is that there exist gaps between samples, even the domain of samples cannot be ensured. Several computational intelligence techniques have been proposed to overcome the limits of learning from small datasets [1], [2], [3], [4], [5].

There are three attribute based approaches for improving classification performance. They are:

1. Attribute selection
2. Feature Extraction
3. Attribute construction

1. Attribute selection

Attribute selection is the process of choosing the subsets of attributes for learning. Feature extraction is the process of turning general representations into more specific ones and attribute construction is creating effective new attributes for knowledge modelling. Since not all the measured variables are important for understanding the underlying phenomena, dimension reduction is often possible in many cases. There may be variables whose variance is less than the measurement noise, and hence these are considered irrelevant to the model.

Conventional methods of feature selection involve evaluating different feature subsets using some indexes and selecting the best among them. The index usually measures the representation capability in the classification or clustering analysis, depending on whether the selection process is supervised or un-supervised. Several techniques have been developed in attribute selection research. [4].

2. Feature extraction

Feature extraction is a technique which has the capability to project the original features into lower feature space to reduce the number of data dimension and improve analytical efficiency. There are two steps that are included in this method the first one is, relevant information for classification is extracted from raw data with original feature vector, m dimension. Second is new feature vector with n dimension (n<m) is created from parameter vector. The methods used for feature extraction include principal component analysis, kernel independent component analysis (KICA), canonical correlation analysis (CCA)and kernel principal component analysis (KPCA).Based on the type of transformation function, feature extraction techniques can be classified into two types: linear and nonlinear. Linear methods, such as principal component analysis, reduce dimensionality by performing linear transformations on the input data and find the globally defined at subspace. These methods are most effective if the input patterns are distributed more or less throughout the subspace. Nonlinear methods, such as KPCA, try to find the locally defined at subspace by nonlinear transformation when the structure of the input data is highly nonlinear[5].

3. Attribute construction

Feature construction is the process of taking a given description of an object and creating a new description of it. Specifically, feature construction refers to the creation of new features which are currently described implicitly by other attributes. One distinction that has been made between attribute construction and feature extraction is that the latter will usually result in significantly fewer features being presented in the data set, while the former adds features to it [6].

In the early research on feature construction, some systems focused on decision-tree-based algorithms. Other methods of attribute construction include genetic based algorithms, and these can be divided into two major categories: the wrapper and non wrapper approaches. For the wrapper genetic programming approach, in which the final learner is used as an indicator for the appropriateness of the constructed attributes, the constructed attributes are fed into the classifier, and the classifier accuracy is used as a guide to rank them. The non wrapper genetic programming approach is performed as a pre-processing phase, and since no particular classifier is involved in evaluating the constructed attributes, it is expected to be more efficient and the results are also expected to be more general. The information gain (IG) and information gain ratio (IGR) are the commonly used fitness functions for constructing attributes.

In Diffusion-neural-network (DNN) method the principle of information diffusion is combined with a traditional neural

 DOI : 10.35291/2454-9150.2021.0437

network, called a diffusion-neural-network (DNN), for functional learning according to the results of their numerical experiments, the DNN improved the accuracy of the Back-propagation Neural Network (BPN). The information diffusion approach partially fills the information gaps caused by data incompleteness via applying fuzzy theories to derive new samples, but the research does not provide clear indications for determining the diffusion functions and diffusion coefficients [7].

The Generalized Trend Diffusion Modelling (GTD) method starts by considering the observations that are collected with an empty set, where the incoming data appear over time. The central location (CL) of data is described and trivially located as the procedure progresses, all the points in a data set appears to be in a batch. GTD is developed to extract information for predicting successive observations. When processing this approach, it generates shadow data employing the real data and the occurrence order of the observed data. Then, it quantities the importance degree for both of the observed and shadow data by computing the membership function values based on fuzzy theories [8].

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

The proposed system architecture is shown in Figure 1. It begins with accepting small samples. Then pre-process the accepted data. After pre-processing build the triangular fuzzy membership function for each class in every attribute. Then on the basis of boundary values of membership function, compute the overlap area for each class in every attribute. Depending on the value of overlap area for each attribute it is set as either low overlap area or high overlap area. If it is low overlap area then class possibility are build otherwise for high overlap area the synthetic attributes are constructed. After that on synthetic attributes apply the Principle component analysis, used to extract the useful attributes. Finally, merge the original attributes, attributes build by the class possibility and the attributes created by attribute construction method. Now this data set is used as learning set for support vector machine to form the classification model.

 DOI : 10.35291/2454-9150.2021.0437

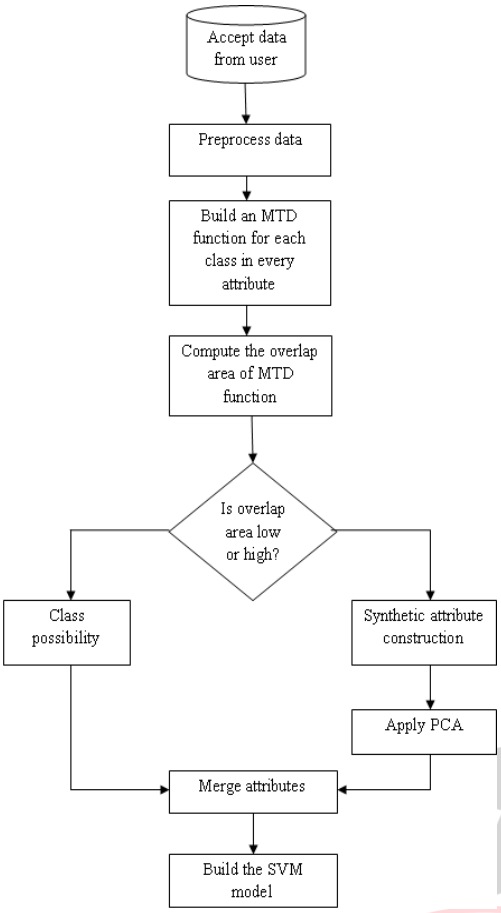

Figure 1: System Overview

The system design includes several main function which are given below.

1. *Building a Mega trend diffusion function for each class in every attribute*

MTD is triangular fuzzy membership function. To solve the problem of insufficient data in small data set analysis MTD was proposed. To calculate the possibility values of virtual samples instead of probability in statistic to avoid the normal distribution assumption, MTD function uses the set theory. Fuzzy membership function is used to substitute the statistical normal distribution to present the probability of sample so as to show that to which class they belong [3].

2. *Pre-process the data*

After accepting the data, it should be pre-processed to remove the noisy data so that the given method is applied to it. If there are some records whose attribute information is missing then the records will be removed or else if the data set contains some categorical information then it will be converted to numeric data by applying some method to it.

3. *Computing the overlap area of each function*

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.

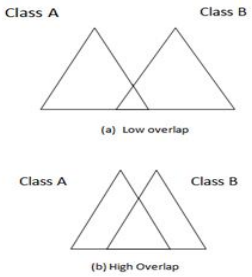

Figure 2 : Low and High overlap

Fig 2 (a) and 2 (b) show the low and high overlap of MTD functions for two classes, A and B, in attribute A1, In attribute A1, the area of overlap of the two classes is low, meaning that attribute A1 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 used to construct new attributes substituting the original ones.

4. *Building class possibility value for low overlap attributes*

This part collects the attributes with low overlap MTD functions and the class-possibility values will be computed to the new attributes to help the judgement of the classes. To Building Up the Fuzzy-Based Transformation, then it is based on the MTD distribution, M(A), considering the classification problem with k-class, transformed A produced by the fuzzy-based transformation.

5. *Constructing synthetic attributes for high overlap attributes*

In this Attribute construction section will discuss the attribute construction process for the attributes 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. Here two methods used one is to compute the correlation matrix and second is combine the highly correlated attributes.

6. *Extracting useful attributes from synthetic method use principle component analysis.*

This method is used for extract the useful attributes from synthetic attribute construction method by applying principle component analysis.

7. *Merging the attributes*

In this, merging the attributes that are build from class possibility method and the attributes that are constructed from synthetic attribute method.

 DOI : 10.35291/2454-9150.2021.0437

8. Building classification model

In this support vector machine classifier is used for building classification model.

IV. MATHEMATICAL MODEL

Let R be the set of dataset, MTD function, Classification model. S = {R,M,C}. We have,

R = {R1,R2,R3...}

M= {M1,M2,M3...}

C = {C1,C2,C3...}

Where R represents the dataset which are input to system, M represents the Mega trend diffusion function that are generated as output of class possibility or synthetic attribute construction and C be the classification model that are classify as output of above two method.

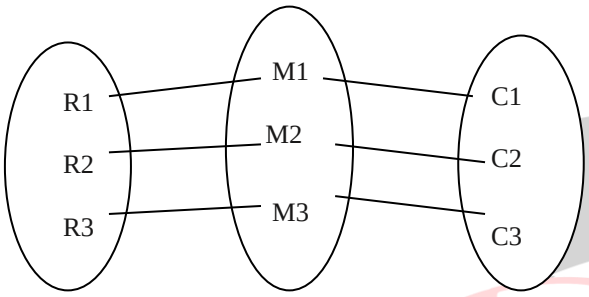

Figure 3: Venn Diagram

Input is mapped to output which is shown in the Venn diagram depicted in Figure 3.

V. RESULTS

The following data set we used to apply this method.

These data set have been downloaded from UCI repository [8]

1. Weather data set: This data set contains weather information. It contains four attributes and one attribute is for class. This is two class problem. It contains numeric as well as categorical data. This data set contains 14 records.

2. Balloon data set: This data set contains weather information. It contains four attributes and one attribute is for class. This is two class problem. This data set contains all attribute information in categorical data. This data set contains 20 records.

Table 1: Sample data set

Records	A1	A2	class
R1	85	85	B
R2	80	90	B
R3	70	96	A
R4	68	80	A
R5	72	95	B
R6	69	70	A

The data set is a sample data set used to generate the method as shown in Table 1. As per the method described above, we have to calculate the boundaries of Mega-trend-Diffusion function.

Table 2: Values of a, U_{set} and b

Attributes	Value of a	Value of U_{set}	Value of b	Class
A1	48.0	78.5	108.15	A
	65.80	69.00	72.19	B
A2	74.00	90.00	105.99	A
	41.04	83.00	124.95	B

Table 3: Class possibility attributes

A1	M_A(A1)	M_B(A1)	A2	M_A(A2)	M_B(A2)
85	438.01	172.21	85	900.00	347.00
80	587.01	105.99	90	831.24	421.47
70	801.16	249.54	96	543.41	410.74
68	744.90	380.00	80	995.11	501.99
72	830.00	453.41	95	954.70	801.90
69	410.45	658.01	70	721.99	658.01

Table 4: Synthetic attributes

Records	A1	A2	A1*A2	A1/A2	A2/A1	class
R1	85	85	7225	1	1	B
R2	80	90	7200	0.88	1.125	B
R3	70	96	6720	0.73	1.37	A
R4	68	80	5440	0.85	1.18	A
R5	72	95	6840	0.76	1.32	B
R6	69	70	4830	0.99	1.01	A

The values for the sample data set are displayed in the Table 2. Table 3 and Table 4 contain the class possibility values and synthetic values. As there are only two attributes available in the sample data set, the correlation is assumed to be high and we have applied the synthetic attribute construction method.

Table 3 and Table 4 will be merged together and will be applied to SVM classifier as the training data set and then the model will be used to predict the class label of unknown data.

VI. CONCLUSION

Classification systems play an important role in business decision-making tasks by classifying the available information based on some criteria. Learning from small data sets is fundamentally difficult, and many data pre-processing methods have been proposed to improve the analysis performance. When the data size is small, there is a high level of uncertainty, and the insufficient often results in less robust analysis. In this paper, thus an attribute construction approach has been discussed which adds data to a small data set and creates a higher dimensional data set which can be used to build classification model. The proposed method can be applied to a data set having

 DOI : 10.35291/2454-9150.2021.0437

multiple class labels. When data set is small it becomes difficult for a classifier to learn from it and even more difficult when there are multiple class labels to approach thus data in small data set is extended by adding more attributes in it which improves classification performance and also avoids problem of over fitting. We have used synthetic data to perform a controlled experiment.

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.

[10] H. Motoda and H. Liu, "Feature Selection, Extraction and Construction," Proc. Sixth Pacific-Asia Conf. Knowledge Discovery and Data Mining, pp. 67-72, 2002

[11] J. Dem_sar, "Statistical Comparisons of Classifiers over Multiple Data Sets," J. Machine Learning Research, vol. 7, pp. 1-30, 2006

[12] Y. Muto and Y. Hamamoto, "Improvement of the Parzen Classifier in Small Training Sample Size Situations," Intelligent Data Analysis, vol. 5, no. 6, pp. 477-490, 2001.

[13] K. Neshatian, M. Zhang, and M. Johnston, Feature Construction and Dimension Reduction Using Genetic Programming, pp. 160-170. Springer-Verlag, 2007.

[14] R. Kohave and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, pp. 273-324, 1997.

[15] C. Kim and C.H. Choi, "A Discriminant Analysis Using Composite Features for Classification Problems," Pattern Recogni-tion, vol. 40, no. 11, pp. 2958-2966, 2007

[16] F.E.B. Otero, M.M.S. Silve, A.A. Freitas, and J.C. Nievola, "Genetic Programming for Attribute Construction in Data Mining," Proc. Genetic Programming: Sixth European Conf. EuroGP, pp. 384-393, 2003.

[17] S. Piramuthu and R.T. Sikora, "Iterative Feature Construction for Improving Inductive Learning Algorithms," Expert Systems with Applications, vol. 36, pp. 3401-3406, 2009.

[18] V. Vapnik, "Universal Learning Technology: Support Vector Machines," NEC J. Advanced Technology, vol. 2, no. 2, pp. 137-144, 2005.

 DOI : 10.35291/2454-9150.2021.0437