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

Dataset Classification by Adding Attribute Information for Improving Classification Accuracy

[1]Yogesh A. Pawar, [2]Prof. Manish Rai

[1,2]Department of Computer Science and Engineering, RKDF C.O.E. BHOPAL (M.P.), India.

Abstract— When we are using minor dataset then main issue is data quantity, because forceful classification operation will not be calculated because of this inadequate or small data. Thus, extracting more efficient information is the new research area which is developing now a days. Considering this new field in this paper we suggest a novel attribute formation method. This method converts original attributes in advanced dimensional characteristic space. This will facilitate to extract additional attribute information using classification oriented fuzzy membership function which is available in similarity-based algorithm. To examine functioning of the suggested method seven datasets having dissimilar attribute sizes are used. From result it is seen that suggested method has efficient classification performance than main component analysis, kernel independent component analysis, and kernel principal component analysis.

Keywords—Inadequate, attribute construction, fuzzy membership function.

I. INTRODUCTION

In this market competition, many situations are there when organizations must work with small datasets. For an example, in the pilot production of a novel product, in the initial stages of a system dealing with a smaller number of dignitary customers, and some unusual cancers, such as for bladder cancer, there are very few medical records available. The important concern in the minor dataset problem is quantity of data, because usually insufficient data will not direct to a robust classification performance. How to extract more effective information from a small dataset is the new research area now [1]. With reference to computational learning theory, sampling size in machine learning problems has a major effect on the learning performance. Looking into this issue, adding some artificial data to the system so as to accelerate acquiring learning stability and to enhance learning accuracy is one effectual approach. In virtual data generation, the prior knowledge is obtained from the minor training set that helps to create the virtual examples for improvement in pattern recognition. Analysts expect to collect more exercise data before learning task is organised, since learning is depended on a small dataset faces the problem of inadequate 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 datasets. Schwarz [4] derived the Schwarz Information Criterion (SIC) using a Bayesian point of view for selection of model, where to choose a model with a major posterior possibility of being correct, Bayes solution is accepted. In machine learning problems, small sample size plays an important role, because without these few sample's information will not be complete. For example, with a classifier, it is problematic to make precise predictions since small datasets not only make the modelling process vulnerable to overfitting but also cause problems in predicting specific correlations between the inputs and outputs. A virtual sample creation approach was proposed for enhancing classification performance for small dataset analysis, but the original concept was suggested by Niyogi et al. [5]. Support vector machines (SVM), are commonly used classifiers.

Today the industrial ecosystem changes rapidly due to globalisation and new inventions. It is notable that the development of products subsequently becomes shorter and shorter. Though the datamining methods are widely used by researchers to obtain correct management info from the data, rare data can be obtained only in the early phases of a manufacture system. From the perspective of machine learning, the amount of training data considerably effects the learning precisions. Learning depended on limited knowledge will be a tough task. Subsequently, at all times investigators want to obtain more training data so as to apply learning tasks; nevertheless, for minor dataset learning, the difficulties encountered critically results from inadequate information.

When learning with small datasets, 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 built on fuzzy model for refining the scheduling knowledge of FMSs, which only have a small dataset for learning. So, in order to full fill the information gaps, a

 DOI : 10.35291/2454-9150.2021.0438

method termed as mega diffusion was replaced by a sample set for diffusing samples one for one. Moreover, a data trend approximation concept is combined with the mega diffusion method to circumvent over-estimating. This method combines mega diffusion and data trend estimation, 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 minimize the number of data dimensions and improve analytical efficiency. These techniques are classified in two types: linear and nonlinear. Linear methods, like main component analysis, reduces the dimensionality by executing linear alterations on the input data. It also discovers generally defined flat subspace. These approaches are most efficient if the input samples are dispersed more or less through the subspace [1].

2.3. Construction of Feature

It is method of creating a new description using existing description of an object. Generally, feature creation 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 dataset [10], while the former adds features to it.

III. PROPOSED SYSTEM

3.1 Building a Mega-Trend Diffusion Function for each Class

The Mega-Trend Diffusion Function was proposed by Li et al. [11] to deal small dataset problem for developing strategies in previous flexible manufacturing systems, and it is a triangular fuzzy membership function. The main reason of the Mega-Trend Diffusion Function is to produce implicit samples to resolve the difficulty of inadequate data in small dataset analysis. To compute the probability values of virtual examples instead of the possibility in numbers to evade the usual distribution hypothesis membership function in fuzzy set theory was used by Li et al.

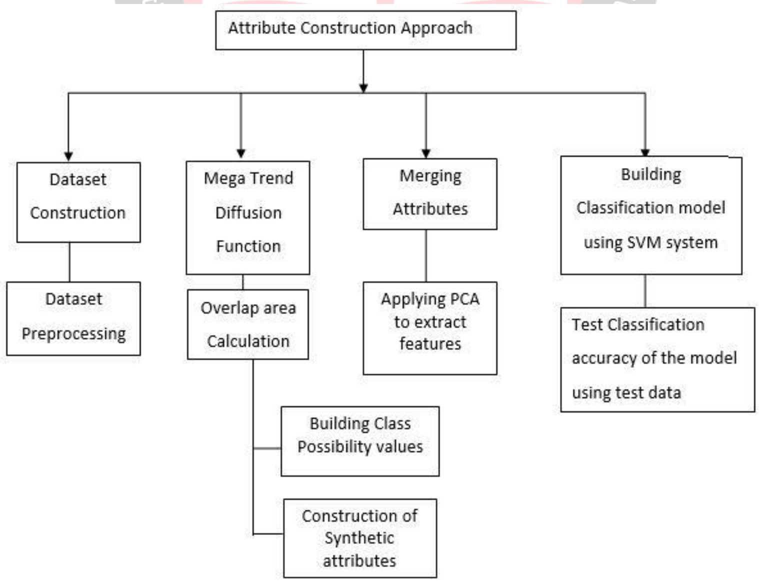


Figure 1. Proposed module breakdown structure

Fig. 2 shows the concept of the fuzzy theorem applied to the Mega-Trend Diffusion 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).

 DOI : 10.35291/2454-9150.2021.0438

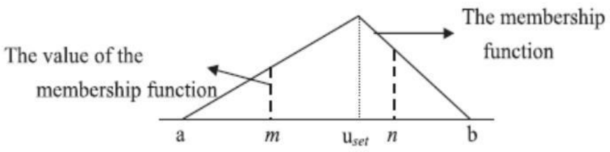


Figure 2. Mega-Trend Diffusion Function

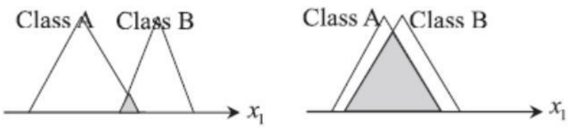


Figure 3. (a) Low overlap (b) High overlap

3.2 Computing the Overlap Area of Mega-Trend Diffusion Function T

After developing the Mega-Trend Diffusion function for every class in each attribute, discovering the overlie part of the Mega-Trend Diffusion functions is an important step for data information extension. Fig. 3 shows an example of Mega-Trend Diffusion function overlapping. Fig 3 (a) and 3 (b) shows the high and low overlap of the Mega-Trend Diffusion functions for two classes Class A and Class B, in the attribute x1, in the attribute x1. The overlapping area of the two classes is small, indicating that the attribute x1 is a useful classification index, since any point in the attribute x1 can be simply classified into the exact 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 dataset to extend the data dimension into a higher feature space to improve the classification precision. For the attributes with a highly overlapping area, the attribute formation technique will be introduced to develop new attributes by 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.

1) Separate the training data into two sets: classes A and B.
2) Apply the Mega-Trend Diffusion technique to compute the membership grade of each attribute in each class.
3) Use the fuzzy-based transformation functions to extend xi into a high dimension.

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

$$\Psi(x) = (x, M^1(x), M^2(x),\ldots\ldots, M^2(x)),\ x \text{ €R}$$

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 $M^1(x)$. The triangle with a dotted line represents the transformation function for class 2, denoted as $M^2(x)$.Thus, for the two-class one-attribute classification problem, the transformed data are $\Psi(x)= (x, M^1(x), M^2(x))$.

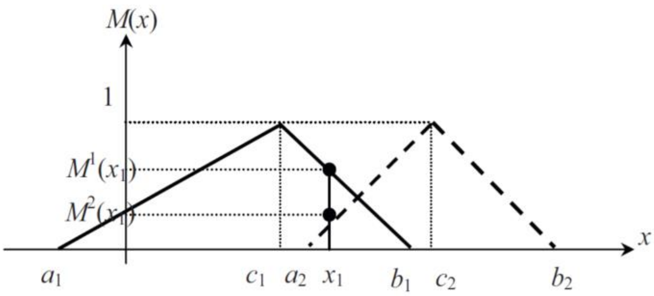


Figure 4.Building up the Fuzzy-Based Transformation Function

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 $X^{\sim k}$ (x1, k), (x2, k),. . . . , k=1,….,K, 0 < S ≤ N.
- Step 2: Starting with class 1, the value of attribute 1, x1, X˜1 of the samples in X˜1 is denoted as x^1i1, i=1,2,…S. Use this value to derive the transformation function for attribute 1, M^1(x1), and repeat this computation for every attribute to obtain M^1(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, M^k (xj), j=1,2, . . .,M , k=1,….,K.

DOI : 10.35291/2454-9150.2021.0438

ISSN : 2454-9150 Vol-07, Issue-05, AUG 2021

- Step 3: Starting with attribute 1, x1, for all samples in X. The transformation function of x1 is set as (x1, M^1(x1), M^2(x1)…., M^k(x)). Repeat this step M times to get (xj, M^1(xj), M^2(xj),….,M^k(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, the computational intricacy of the fuzzy-based alteration that transform the original dataset into the new space is O(NxKM).
- Step 4: After all samples have been altered, PCA is used to extract the features. In this, the computational complexity of estimating the PCA is

$$O((M \times (K+1))^2 N) + O((M \times (K+1))^3)$$

Step 5: Input the dataset 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 $nSV + nSV^3$), where *nSV* denotes number of support vectors for a problem.

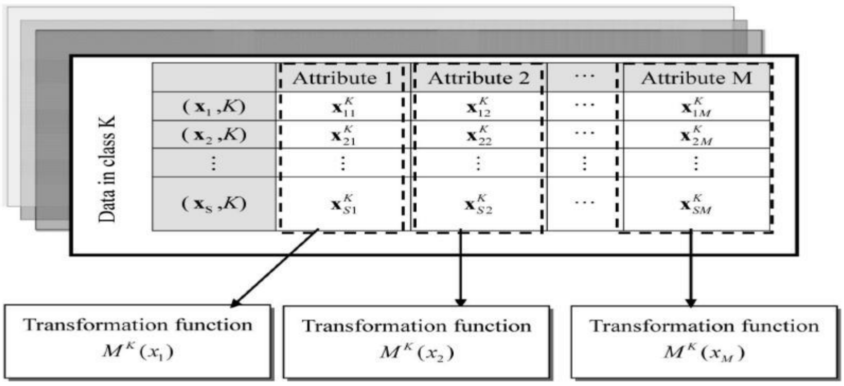


Figure 5. Transformation function building process for each attribute of samples in class 1.

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. The Pearson correlation coefficient is defined as

$$\Sigma(\; , \;) = \sqrt{\frac{(\; , \;)}{(\;) + (\;)}}$$

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 A^2. 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 A^2 is also not considered here. In short, this paper collects the attributes with a high overlap area of Mega-Trend Diffusion function, and then computes the correlation matrix. Any pair of attributes, A and B, with a high correlation value will be used to construct the new attributes which are A*B, A/B, and B/A. Finally, PCA will be used to extract the features by using attributes A, B, A*B, A/B, and B/A.

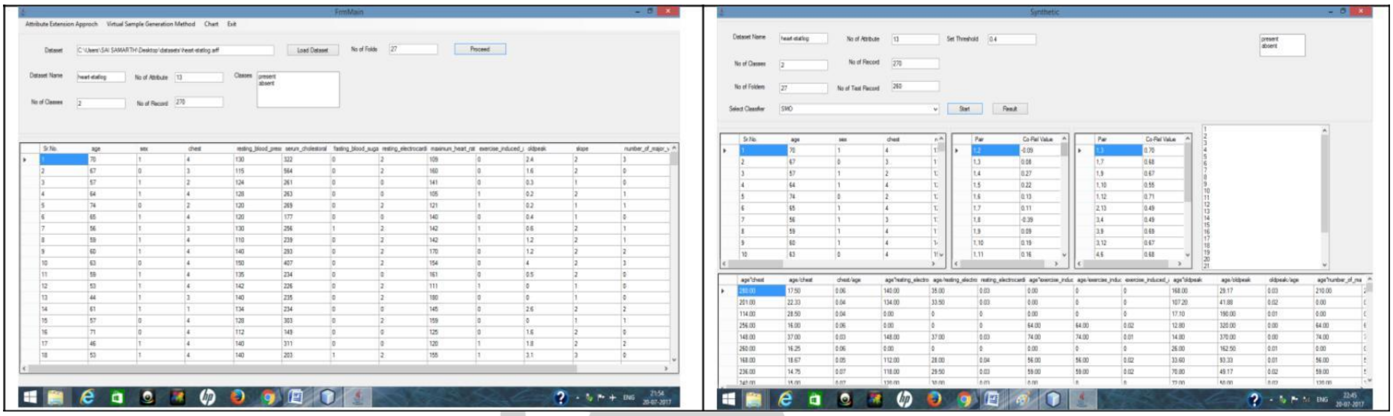
 DOI : 10.35291/2454-9150.2021.0438

3.5 Build SVM Model

SVM classifier with a Gaussian kernel is used for building a classification model after preprocessing dataset 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 dataset. User can select the data pre-processing method from the menu bar. After loading the dataset and selecting the method, a classifier is chosen. After entering the size of each fold, the dataset 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 dataset 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

After observing Table 1, it is seen that real small dataset is not avail-able with this paper except bladder cancer dataset. Therefore the dataset is divided into number of folds, each fold of less number of records is used as training dataset and rest of the folds are used as test data. This procedure is repeated for all the folds and then the average accuracy is computed.

Attributes Extended for each dataset with attribute extension method by generating only class possibility values and only synthetic attributes are presented in above Table 2.

 DOI : 10.35291/2454-9150.2021.0438

The correlation coefficient for bladder cancer dataset is displayed in below table 3.

((‾)
,
=1 − =1 , −

Table 3. Accuracy of Heart Statlog dataset for 10 training samples(27 fold)

Method used	SMO	J48	Naivebayes	Logistic	Multilayer Perception
Original	73.54	69.52	69.95	69.31	71.50
Only class possibility	71.15	68.56	67.63	69.56	72.33
Only synthetic	69.85	69.22	67.44	69.81	70.85
Attribute construction approach combined	67.13	65.56	61.66	65.54	70.85

It is observed that in the Heart stat-log dataset accuracy is better in original dataset compared with all preprocessing methods, this is shown in table 3.

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 dataset. 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", 452-464, 2012.

[2] Shawe-Taylor, M. Anthony, and N.L. Biggs, "Bounding Sample Size with the Vapnik-Chervonenkis

[3] Dimension," Discrete Applied Math., vol. 42, pp. 65-73, 1993.

[4] Y. Muto and Y. Hamamoto, "Improvement of the Parzen Classifier in Small Training Sample Size Situations,"

[5] Intelligent Data Analysis, vol. 5, no. 6, pp. 477-490, 2001.

[6] G. Schwarz, "Estimating the Dimensions of a Model," Analysis of Statistics, vol. 6, pp. 461-464, 1978.

[7] P. Niyogi, F. Girosi, and P. Tomaso, "Incorporating Prior Information in Machine Learning by Creating Virtual

[8] Examples," Proc. IEEE, vol. 86, no. 11, pp. 2196-2209, Nov. 1998.

[9] M.A. Hall and G. Holmes, "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining," IEEE

[10] Trans. Knowledge and Data Eng., vol. 15, no. 6, pp. 1437-1447, Nov./Dec. 2002.

[11] R. Thawonmas and S. Abe, "A Novel Approach to Feature Selection Based on Analysis of Class Regions," IEEE

[12] Trans. Systems, Man, and Cybernetics, Part B: Cybernetics, vol. 27, no. 2, pp. 196-207, Apr. 1997.

[13] M. Muharram and G.D. Smith, "Evolutionary Constructive Induction," IEEE Trans. Knowledge and Data Eng., vol. 17, no. 11, pp. 1518-1528, Nov. 2005.

[14] P. Mitra, "Unsupervised Feature Selection Using Feature Similarity," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301-312, Mar. 2002.

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

[16] V.S.Verykios, A.K. Elmagarmid, E. Bertino,Y. Saygın, and E. Dasseni, Association Rule Hiding, Jan 7, 2003.

 DOI : 10.35291/2454-9150.2021.0438