An Implementation of Data Pre-Processing for Small Dataset

Sameer Dixit
Senior Assistant Professor,
Department of Computer Science & Engineering,
Malwa Institute of Technology, Indore,
Madhya Pradesh, India.

Navjot Gwal
Department of Computer Science & Engineering,
SDBCT, Indore,
Madhya Pradesh, India

ABSTRACT
Pre-processing in data mining played essential role for enhancing data quality. The basic concept behind is that, learning with accurate and high quality data may provide more efficient classification results as compared to learning with poor quality of data. In this presented paper a pre-processing technique is implemented with slight modification which is based on the technique given in [1]. In this paper a promising approach of data pre-processing is provided which utilizes a fuzzy technique in order to improve the data quality. The implementation of available technique is performed using MATLAB. Additionally, the improved fuzzy technique is also implemented with it. The results demonstrate the effectiveness in classification accuracy after implementation of both techniques. Finally, the obtained results favour the proposed model for enhancing the performance of classifiers in both manners supervised and unsupervised manner.

Keywords—data mining,pre-processing, data quality enhancement, classification, performance improvement

1. INTRODUCTION

Data mining is an approach to find the meaningful patterns from data.This meaningful content may helpful for decision making, classification, large scale data analysis and other similar intelligent task.In data mining the main and basic element is data. Mining of data and information recovery is directly depends upon data. Therefore, learning process of a data mining algorithm is majorly depends upon the type of data and quality of data.For example, a student can better learn with quality of knowledge delivered by a book, but some of the pages missing in book can effect in students learning. Therefore, a data mining algorithm required high quality of data for effective learning.

In data mining process, learning can be categorized as supervised and unsupervised learning technique. In supervised learning a trainer is available, mean to say the training data includes the attributes and their outcomes. On the other hand in unsupervised classification the data contains only attributes there are not any class labels exist.

The data mining process may works as process stack, where after completion of a process another process take place. The three basic processes are resided in this process stack, pre-processing, learning and finally the validation of learning or testing.

In this presented paper the main focus is to work with pre-processing phase that can help in classification performance improvement. In this section a general, overview of the work isprovided, the next section includes the background of the studying domain.

2. BACKGROUND

In this section, the background work and similar methodology is discussed.Additionally, the basics of the implemented module are discussing using this section.

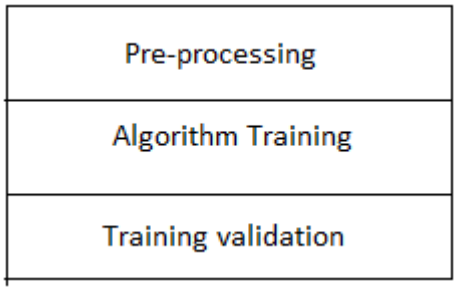

Figure 1 data mining steps

As given figure 1 the pre-processing is first process of the data mining process. Data quality control is the process of controlling the usage of data with known quality measurement—for an application or a process. Data Quality Assurance (QA) process consists of discovery of data inconsistency and correction. Data QA process provides following information to Data Quality Control (QC):

- Severity of inconsistency
- Incompleteness
- Accuracy
- Precision
- Missing / Unknown

After creating data files an indispensable key step in data analysis processes is to prepare data for further analysis. In fact, according to a study data pre-processing occupies about 70% of time spent on knowledge discovery project. If data have impurities, for example missing or duplicate data, data mining tools may be misled and even give wrong results. Based on wrong results, companies may make fatal decisions. Besides, preparing data is an integral part of building a data warehouse so that it integrates data of uniform quality [2].

A dataset is a collection of data that can be representable in tabular form. In tabular manner columns represents a feature, more frequently called attributes. These are may be in form of continuous, categorical, or binary. In such data tables' rows represents a correlation between these features or attributes, in terms of data mining that are known as instances. According to the amount of instances available in a dataset, that can be defined by two types, large datasets and small datasets. In large datasets, the number of features is more for prediction accurately. In small datasets, the number of features is less as compared to large datasets and amount of data instances are also too few. Therefore small dataset contains less information, and not able to define the actual scenarios of application problem.

Therefore a dataset with low dimensions and less amount of meaningful contains are not much suitable for train a data model, results classification performance of a classifier decreases significant amount. Therefore, feature construction methods are used to generate more relevant features of datasets. Feature construction consists of constructing new features by applying some operations or functions to the original features, so that the new features make the learning task easier for a data mining algorithm. The classification accuracy of a classifier can often be significantly improved by constructing new features which are more relevant for predicting the class of an object [3].

This section presents the basic general discussion of the pre-processing and there effect over data set, the next section provides the existing technique of pre-processing that required to modify.

3. EXISTING WORK

If the datasets contains less number of features that affect the performance of a classifier. Therefore, feature construction method may helpful for improving datasets. In order to enhance features of dataset, fuzzy membership function known as Mega-Trend Diffusion (MTD) is used to build features [1]. After that the overlap area of the Mega-Trend Diffusion for each class is required to compute. If computed area of membership function is less, then class-possibility building method used. On the other hand if the overlapping area of is high then synthetic feature construction is used.

During synthetic feature construction correlation coefficient is computed. Each pair of features (A, B) with a high correlation relation can used to create the new features.

For feature extraction from the new datasets obtained by above method Principal Component Analysis (PCA) is used as a feature extraction technique, that also helps to reduce the dimensionalityfinally, required merging features obtained by above method.

The complete process as defined in [1] can be summarized as:

Assuming that we have sample set$X = \{x_1, x_2, \dots, x_N\}$, where each sample$x_i, i = 1,2, \dots, N$ in X has M attributes (means$x_i = (x_1, \dots, x_M)$).

1. Compute the correlation coefficient matrix from the data set X.
2. Extract the pairs of attributes, (y_i, y_j), for which the correlation value is larger than 0.4 (the lower bound of moderate correlation).
3. Construct the synthetic attributes using the operators * and / For example, if the correlation value of the pair of attributes (y_i, y_j) is larger then 0.4, then the corresponding synthetic attributes are defined as y_i y_j;y_i=y_j, and y_j=y_i.
4. After all the synthetic attributes are built, the PCA is used to extract and ortho-normalize the features.

The above defined model is implemented using MATLAB, additionally in order to enhance performance of the classifier the next section describe the work.

4. PROPOSED WORK

This section of the paper includes the proposed improvement on traditional method [1]. The first modification is in place of PCA (principle component analysis) the use of KPCA (kernel principal component analysis) is used.

Kernel Principal Component Analysis

Kernel Principal Component Analysis (KPCA) is an extension of PCA using techniques of kernel methods. KPCA has proven to be a powerful tool as a nonlinear feature extractor of classification algorithm. The basic idea is to map the data from the input space to a feature space F through some nonlinear function then applying the linear PCA there. Through the use of a suitably chosen kernel function, the data becomes more linearly related in the feature space F.

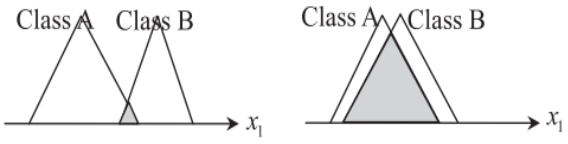

Figure 2 high and low overlapping

Second modification on the computation in "Computing the Overlap Area of MTD Function", the overlap area is given by figure 2.

In [1] thehighness of the overlap area is computed using the idea ofthe geometric mean. For example, given a two-class, A andB, data set X, the area of the MTD function of attribute i ofclass A in X isβ_A^i, the area of the MTD function of attribute I of class B in X is β_B^i, and the overlap area of two MTDfunctions of classesAandBisβ_O^i. The rate of overlap of class A is β_O^i/β_A^i, and the rate of overlap area of class B isβ_O^i/β_B^i.The overlap degree in attribute i can thus be defined as

$$OD^i = \sqrt{\frac{\beta_O^i}{\beta_A^i} \cdot \frac{\beta_O^i}{\beta_B^i}}$$

The threshold of OD^ias the average ofOD, and the corresponding attributes are defined as havinglow overlap when less than OD, and otherwise as havinghigh overlap.

$$\begin{cases} OD^i < mean\ (OD), & Low\ overlap \\ OD^i \geq mean(OD) & High\ overlap \end{cases}$$

Where OD is a vector for which the components areOD^i for$i = 1, 2, \dots, M(OD = (OD^1, OD^2, \dots, OD^i))$there are tworeasons to use the average value of OD. One is thethreshold is dynamic, and the other is to prevent allattributes belonging to low or high overlap groups.

But, to get more precise values for creating a threshold required improving the threshold in place of computing mean values, use with the precision values. That may help to improve the computation of high and low overlap area.

The next section provides the comparative outcomes as experimental results.

5. RESULTS

The proposed improvement on the MDT function for construction optimum feature set is evaluated through the presence of both kind of classifiers supervised (SVM) and unsupervised (KNN) algorithm. For experimental evaluation the machine learning datasets given in [5] is consumed and performance is evaluated and provided below.

Using supervised classifier: to demonstrating the effectiveness the comparative results using SVM classification algorithm in three different scenarios is provided. In first original data is used for classification, secondly using the given method in [1] the performance of classifier evaluated and finally using proposed modifications the outcomes evaluated, The outcomes of classifier is given using figure 3 and 4.

In the given figure 3 the data sets that exactly contains two classes, the effect of classification is provided. In addition of that for multi-class (more than two classes) the performance of classifier is given using figure 4. In two class data set performance of SVM classifier is improved by both the techniques. On the other hand for the multiclass datasets the performance of classifier is not much enhanced by the method given in [1]. But the proposed modified method is capable to enhance the classification accuracy in both the conditions.

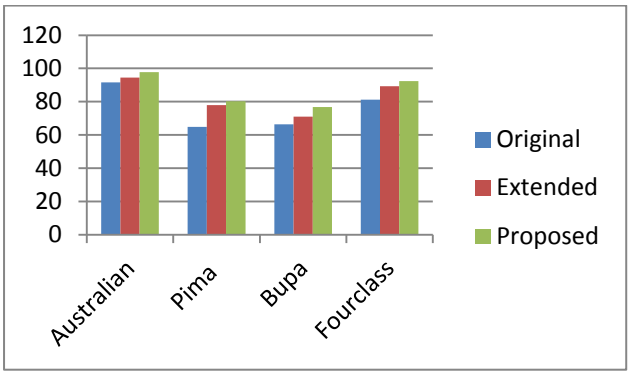

Figure 3 two class classification

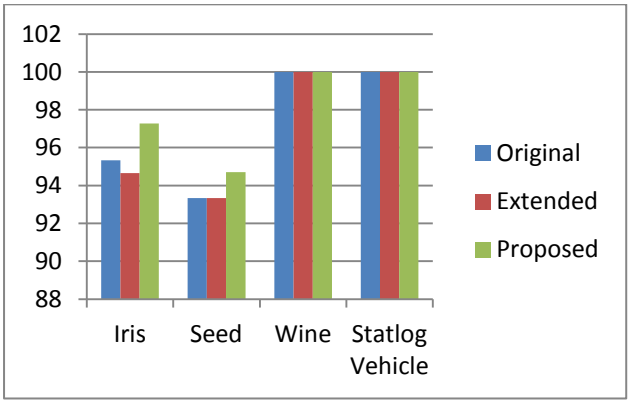

Figure 4 multi class classifications

Using unsupervised classifier: the figure 5 and 6 represents the classification performance of both unsupervised classifier namely KNN.

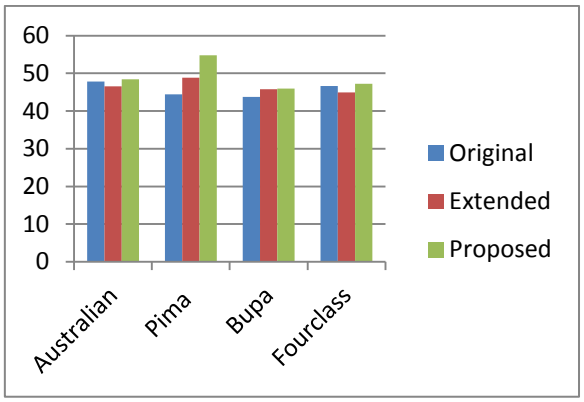

Figure 5 two class classification

According to the figure 5 the performance of unsupervised classifier is not much accurate but both the method improves the performance for two class classification datasets. On the other hand the improvement is much optimized for multi-class datasets as the results obtained results from KNN classifier as given in figure6.

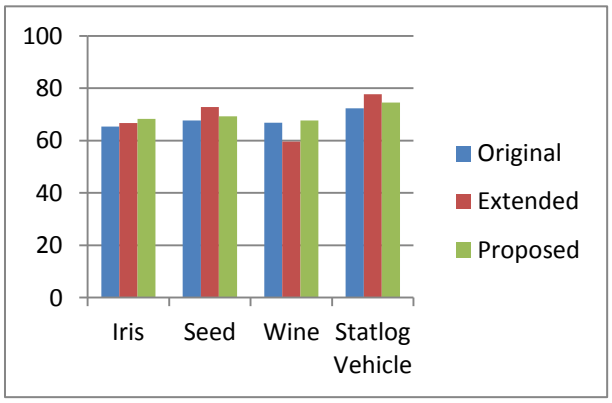

Figure 4 multi class classifications

This section provides the classification results obtained in both kind of classifiers. The next section concluded the complete work performed in order to improve the data quality.

6. CONCLUSIONS

This paper is intended to study about the pre-processing method in data mining techniques.In this direction various essential contributions are found for improving the dataset quality. And that is found that for efficient learning process the high quality of data is required. For that purpose a method described in [1] is studied and implemented. During the implementation, some additional experiments are performed for improving the existing method. After implementation of both available in [1] and proposed method that is found both of the methods helps in improving the classifiers performance. The experimental results demonstrate the effectiveness of the given method for both kinds ofclassifiers (supervised and unsupervised). Additionally that is also find that the improvement in classification for two classes are much effective than the multi-class datasets.

7. REFERENCES

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

[2] Yifei Ren, "Data Preprocessing for Data Mining", Bachelor's Thesis (UAS) Degree Program in Information Technology Information Technology 2013.

[3] Rayner Alfred, "Optimizing Feature Construction Process for Dynamic aggregation of Relational Features," Journal of Computer Science, 2009 Science Publication.

[4] X. Sun, S.Z. Sun, J. Tian and J. Han, "Sparse Kernel Principal Component Analysis on Seismic Denoising and Fluid Identification", 10 June 2013 DOI: 10.3997/2214-4609.20130642

[5] Index of /Datasets/UCI/arff - Parent Directory – Seasr: http://repository.seasr.org/Datasets/UCI/arff/