© 2015 IJSRSET | Volume 1 | Issue 2 | Print ISSN : 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

A survey on Improving Classification Accuracy in Data Mining

Bhavesh Patankar*[1], Dr. Vijay Chavda[2]

*[1]Department of M.Sc. (IT), Kadi SarvaVishwaVidyalaya, Gandhinagar, Gujarat, India.

[2]NPCCSM, Kadi SarvaVishwaVidyalaya, Gandhinagar, Gujarat, India.

ABSTRACT

There are various classifiers available for data classification, selecting the best classifier is one of the critical problems of data classification. Also pre-processing approach to be used is quite important. In this paper, study of various approaches to improve the classification accuracy in data mining is carried out. The purpose of the pre-processing is to gain a high degree of distinct classes before the classifier is trained or tested. Handling noise and outliers is an important aspect in data mining to improve the classification accuracy. High accuracy of classification also depends upon the quality of data being used for classification in data mining. Feature selection is also one of the aspects which can refine the dataset before providing it to the learning algorithm to improve the accuracy of the classifier.

Keywords: Classification; Pre-processing; Outliers detection; Feature Selection; Dimensionality reduction

I. INTRODUCTION

Finding the accurate classification of the data is the main goal of classification in data mining. Owing to this, different approaches have been suggested to improve the accuracy of the classification. One of the approaches relies on the data and tries to improve classification accuracy through using pre-processing techniques before performing classification. During the past decades, various techniques have been suggested for pre-processing which includes feature selection, dimensionality reduction, noise reduction, etc [1].

In this paper we have reviewed various papers to analyze the different parameters to be considered in order to improve the classification accuracy in data mining. A pre-processing step is recommended before the classification is performed in order to achieve increasing discrimination of classes. The available data set is transformed into more qualitative data set. Sometimes it may happen that dataset contains high dimensions, some of the dimensions may be irrelevant for our classification purposes. So it may be required to perform Feature selection to utilize the best features for achieving the higher accuracy in classification. Various methods have been suggested to overcome noise and outliers in order to improve the classification accuracy.

II. FEATURE SELECTION

Higher accuracy is vital in any data mining process. Feature selection aims in selecting a subset of relevant features for building robust learning models. Camelia Vidrighin et al.[2] has focused on combining feature selection with missing values citation to improve the performance of the learning methods. They have analyzed the various approaches for feature selection, and identified those which yielded considerable accuracy improvements. Nilsson[3] concluded that there exist two main types of feature selection problems: "(i) finding the predictive features, which are important for building accurate predictors, and (ii) finding all features relevant to the target variable". In machine learning, Feature selection algorithms are traditionally divided into filter methods (filters) and wrapper methods (wrappers). Filter methods are being motivated by the properties of the data division itself. There are various robust algorithms in literature which utilizes a filter strategy. The filter model relies on general characteristics of the training data to select some

features without involving any learning algorithm, whereas the wrapper model requires one predetermined learning algorithm in feature selection and uses its performance to evaluate and determine which features are chosen. As for each new subset of features, the wrapper model needs to study a hypothesis (or a classifier). It tends to find features better suited to the predetermined learning algorithm resulting in superior learning performance, but it also suppose to be computationally more expensive than the filter model.

Feature selection is an important step in a data mining process. Among the many possible advantages of feature selection, we are particularly interested in improving the classification performance. Camelia Vidrighin Bratu et. al[2] have considered the wrapper approach, as a combination of three steps: generation, evaluation and validation. The main focus is to analyze the behavior of the most promising combinations of search methods for generation, and inducers for evaluation and validation, such that the performance (i.e. the accuracy) increase is guaranteed. Just like in the case of learning algorithms, there is no universally best feature selection method, i.e. a method which yields the highest accuracy improvements on any dataset, over the other methods. However, as presented by Kothavi [4], for the purpose of accuracy improvement wrappers seem to be the most appropriate approach. In order to provide a comprehensive analysis of the chosen methodology, we have initially considered four different generation procedures (forward greedy stepwise, backward greedy stepwise, forward best-first search and backward best first search). Initial results by Vidrighin et al. [5] have shown that two search methods perform significantly better than the others: backward greedy stepwise and forward best-first search.

Experimental results given by Camelia Vidrighin Bratu et. al[2] prove that wrappers always improve the performance of classifiers. In many cases, the inducer which initially achieved the highest accuracy maintains its high performance after feature selection (first or second best performance). This means that once we have initially assessed the dataset and selected a learning scheme as being appropriate, that scheme will maintain its performance all the way through the data mining procedure.

Also, for all the datasets considered, the second best performance after feature selection still yields significant improvements over the initial inducer, which proves the necessity for such a step.

III. DATA CLEANING

Existence of outliers and noise is quite common in dataset due to errors like typographical errors. This outliers and noise can cause the model to be built is weak model. To make the better model it is vital to improve how learning algorithm can handle the noise and outliers. Success of any data mining problem depends upon the quality of the data. It is in general acknowledged that the data preprocessing step is necessary for obtaining a healthy data mining process. Among the preprocessing activities, missing values imputation and feature selection are the very important. Removing entities that are noise is an important goal of data cleaning, as noise hinders most types of data mining activities. Almost all existing data cleaning methods concentrate on removing noise that is the result of low-level data errors that result from an defective data collection process, but data objects that are immaterial or only weakly relevant can also significantly hamper data analysis. So, if the goal is to improve the data analysis as good as possible, these entities should also be measured as noise, at least with respect to the underlying data mining process. Accordingly, there is a need for data cleaning methods that remove both types of noise. As the data sets can contain large amounts of noise, these methods also need to be able to remove a potentially large fraction of the data. Raw data is highly vulnerable to noise, missing values, and inconsistency. The data quality affects the data mining output. In order to improve the quality of the data and, accordingly, of the data mining results raw data is pre-processed so as to improve the efficiency and ease of the mining process. Data pre-processing is one of the most important steps in a data mining process which deals with the preparation and conversion of the initial dataset. Data Cleaning is one of the techniques of Data pre-processing activity.

Data that is to be consumed by data mining techniques can be imperfect (missing attribute values or certain attributes of interest, or containing only collective data), erroneous (containing errors, or outlier values which is not match with the expected), and inconsistent (e.g., containing discrepancies in the category codes used to

categorize products). Noisy, incomplete and inconsistent data are now the part of huge databases and data warehouses. Incomplete data can occur for a number of possible reasons. Sometimes important attributes might not available, e.g. Supplier information for purchase transaction data. Some data may be ignored, may be because it was not considered valuable at the time of data entry. Appropriate data may not be recorded due to a misunderstanding, or because of tools malfunctions. It might be possible that data inconsistency with other recorded data may have been deleted. Adding to it, the recording of the history or modifications to the data may have been ignored. Omitted data, mainly for tuples with absent values for some attributes, may need to be filled. Dataset can be noisy may be due to following reasons. The tool used to collect the data may be erroneous. There may be human or computer errors occurring at data entry. In data transmission, error may occur during the process. Inaccurate data may also arise from variation in naming conventions or data codification used. To make the consistent data, duplicate data also need to be cleaned. Data cleaning process clean the data by way of getting in missing data, correcting the noisy data and finding out the outliers and removing it, and resolving inconsistencies. Dirty data can make uncertainty for the data mining practice. Even though most of the mining techniques have some procedures for dealing with noisy or incomplete data, they are not always strong. As an alternative, they may concentrate on avoiding over fitting the data to the function being modeled. So, an important pre-processing step is to do the data cleaning activities on your data.

Missing Values: During analysis, if it is observed that there are many entities that have no recorded value for numerous attributes, then those missing values can be put in for the given attribute by a variety of methods as given below:

1. Ignore the tuple: This method is usually used when the class label is missing (assuming the data mining task involves classification). This method is very useful, when the tuple contains several attributes having missing values. When the percentage of missing values per attribute varies remarkable, it shows weak performance.
2. Manually filling the missing values: This method is time-consuming and may not be practical given a large data set with numerous missing values.
3. Utilising a global constant to fill in the missing value: Replace all missing attribute values by using constant like "missingtext". It may happen that if we replace the missing values by "missingtext", then the mining program may mistakenly think that they form an interesting pattern as they all have common value of "missingtext". Even though this method is simple, it is not recommended to use.
4. We can calculate the mean of the attribute and use it to fill the missing values.
5. To fill the missing values for the samples belonging to the same class, we've to use the attribute mean of those samples.
6. We can use the various tools like decision tree induction etc. to find out the most probable value for the missing value to fill it. Using method 3 to 6, it may happen the filled-in value may not be true. However, method 6 is very popular in compare with the other methods as it uses all the information from the present data to predict missing values.

Noisy Data: Noise refers to the modification in the original value. We need to use the technique to smooth the data to remove the noise.

The following data smoothing techniques describes this

1. Binning methods: Binning methods smooth data value by referring the value of its neighbour or value around it. In this method sorted data is separated into equal number of bins or buckets then smoothing can be done by bin median or by bin end points. Let us have sorted data for item price: 5, 9, 16, 22, 22, 25, 26, 29, 35.

Now splitting the data into equal groups.
Bin1: 5,9,16
Bin2: 22,22,25
Bin3: 26,29,35

Smoothing by bin means:
Bin1: 10,10,10
Bin2: 23,23,23
Bin 3: 30, 30, 30
Smoothing by bin boundaries:
Bin1: 5,5,16
Bin2: 22,22,25
Bin 3: 26, 26, 35

2. Regression: Data can also be smoothed by using regression methods. Linear regression can be used to

find the best line between the two points. One point can be used to predict the other. Multiple linear regression can also be used which is an extension of linear regression. In multiple linear regression more than two variables are concerned and the data are fit to a multidimensional plane. Using regression, mathematical equation is derived which is used to smooth the noisy data.

3. Outlier Detection: Outlier can be detected using clustering. When similar data values are organized into groups, so that the value which doesn't fall under any cluster or group can be easily identified as outlier. [8]

Outlier detection has acknowledged increasing awareness, mainly from the data mining perspective where outliers may correspond to anomalies or center of interest. One of the prominent problems in outlier detection is that there are no established criteria of what makes an outlier. Considering broad-spectrum, outlier detection methods have used artificial data sets or have injected noisy instances into a data set to set up which instances are outliers. Also, there are many outlier detection algorithms from a variety of fields using different approaches; a few techniques are reviewed here. Khoshgoftaar et al [6] used a outlier detection method to remove outliers. They have done analysis of their approach by synthetically injecting noise into clean data from software measurement data of a NASA software project. Michael R. Smith et al[7] used PRISM (PReprocessing Instances that Should be Misclassified), a novel filtering method that identifies instances that should be misclassified. In PRISM, they used 3 outlier detection approaches and 1 noise reduction method to train 9 learning algorithms with filtering and compared the results to those from the learning algorithms trained using the original data set. They achieved on average, the increase in accuracy was about 1.3%.

IV. CONCLUSION

In this paper, it is found that classification accuracy can be improved by various approaches. Feature selection is one of the approaches to improve the classification accuracy. It is important to select the best features while classification from the dataset, because it might happen that sometimes one of the important feature get removed during this exercise, which causes improper classification. Data cleaning is also very important activity to perform in order to improve the classification accuracy. Data cleaning includes filling the missing values, outlier detection, correcting the noisy data etc. Using above all techniques definitely improve the classification accuracy.

V. REFERENCES

[1]. Moeinzadeh, H, Nasersharif, B, Rezaee, A., Pazhoumand-dar, H., "Improving Classification Accuracy Using Evolutionary Fuzzy Transformation", 11th Annual Conference on Genetic and Evolutionary Computation Conference (GECCO 2009), Montreal, Canada, 2009 (1)

[2]. Bratu, C.V.; Muresan, T.; Potolea, R., "Improving classification accuracy through feature selection," Intelligent Computer Communication and Processing, 2008. ICCP 2008. 4th International Conference on , vol., no., pp.25,32, 28-30 Aug. 2008.

[3]. Nilsson,R., Statistical Feature Selection, with Applications in Life Science, PhD Thesis, Linkoping University, 2007.

[4]. University, Kohavi, R. Wrappers for Performance Enhancement and Oblivious Decision Graphs, PhD thesis, Stanford University, Computer Science Department, 1995. (3).

[5]. Vidrighin C., Potolea R., „Towards a Combined Approach for Feature Selection", accepted at ICSOFT 2008.

[6]. T. M. Khoshgoftaar, N. Seliya, and K. Gao. Rule-based noise detection for software measurement data. In Proc.of the IEEE int. conf. on inf. Reuse and Integration,pages 302–307. IEEE Syst., Man, and Cybern. Society, 2004.

[7]. Smith, Michael R., and Tony Martinez. "Improving classification accuracy by identifying and removing instances that should be misclassified." Neural Networks (IJCNN), The 2011 International Joint Conference on. IEEE, 2011.

[8]. Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining, southeast asia edition: Concepts and techniques*. Morgan kaufmann, 2006.