IJETST- Vol.||02||Issue||10||Pages 3271-3276||October||ISSN 2348-9480 2015

Open access Journal International Journal of Emerging Trends in Science and Technology

DOI: http://dx.doi.org/10.18535/ijetst/v2i10.06

Refining the Noisy Candidates in Learning Data for Improving Classification Performance

Authors

Maya Yadav[1], Shubhangi Sharma[2]

[1]Assistant Professor Sanghvi Institute of Management and Science Indore, India

[2]M.Tech Research Scholar Sanghvi Institute of Management and Science Indore, India

Abstract

The data mining is a technique are used to learn a pattern and identify the pattern among a huge amount of data using computer based algorithms or programs. In this technique the two main technique of learning is found first supervised and second the unsupervised. But the quality of learning is depends upon the amount of data and quality of learning pattern. Therefore, in order to improve the quality of learning data the pre-processing is performed on data. In this paper a review on the pre-processing techniques and the data quality enhancement techniques is reported. Further the key issues and challenges are addressed for improving the learning data quality. Finally for resolving the issues a new technique is proposed in this paper and their future extension of the work is also provided.

Keywords: *data mining, pre-processing, learning, supervised techniques, data quality assessment*

Introduction:-

There is a huge amount of data is available in information industry. If we want to utilize this data for increase revenue and decrease cost so we have to use data mining. Data mining is a basically an extraction and analysis of large quantity of data from which we can discover valid, novel, potentially useful and ultimately understandable pattern in data [1]. Data mining can be used in different purpose and prospective. With the help of data mining we can predict behaviour, future aspects, facts, trends, valuable data, hidden relationships and anomalies and increase data access [2]. Data mining is also known as Knowledge Discovery in Data (KDD).Data mining is not only extract data from data base but it's also including data integration, data selection, data cleaning, data transformation etc.

Data integration:- all the data is collected and integrated from different sources[3].

Data selection:- Identification of target data sets and relevant fields[3].

Data cleaning:- The data which is collected may be not clean , included noise ,missing errors so we have to clean data remove noise and outliers, create common units, generates new fields[3] .

Data transformation:-Transfer data in to appropriate formats. For data transformation we use different techniques smoothing, aggregation, normalization etc. [3].

There are some applications of data mining:-

- Design and construction of data warehouses for multidimensional data.
- Classification and clustering of customers for targeted marketing.
- Multidimensional analysis of sales, customers, products, time and region.
- Identification of unusual patterns.
- Association and correlation analysis, aggregation to help select and build discriminating attributes.

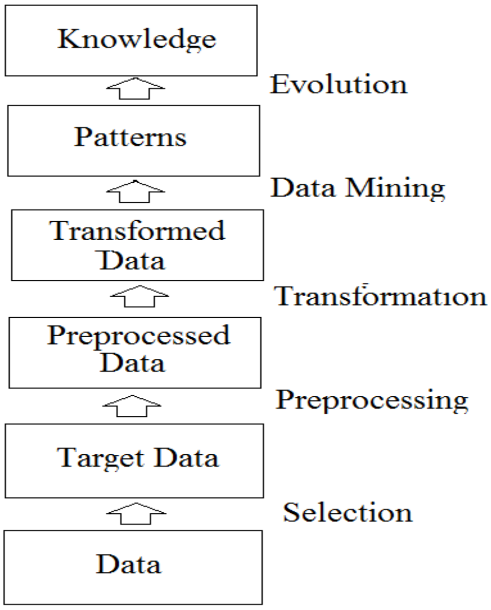

Fig.1 pre-processing method

Fig.1shows the pre-processing method in which data is collected from the different sources this is called data collection. After collection of data we select the data on which we want to perform pre-processing methods this is called target data .on target data we apply different methods of pre-processing such as integration, selection, cleaning etc. Pre-processed data is transform into that form in which that is required. From these transformed data we find patterns. Patterns generation from the raw data is called data pre-processing.

Background:

Data pre-processing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviours or trends, and is likely to contain many errors [4]. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. Data pre-processing transforms the data into a format that will be more easily and effectively processed for the purpose of the user.

Steps of data pre-processing:

1) Data cleaning: -

Data cleaning is a technique that is applied to remove the noisy data and correct the inconsistencies in data [4]. Data cleaning involves transformations to correct the wrong data. Data cleaning is performed as a data pre-processing step while preparing the data for a data warehouse. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute [5]. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data [6].

2) Data transformation: -

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system. The usual process involves converting documents, but data conversions sometimes involve the conversion of a program from one computer language to another to enable the program to run on a different platform [7]. Data transformation can be divided into two steps:

- Data mapping maps data elements from the source data system to the destination data system and captures any transformation that must occur [7].
- Code generation that creates the actual transformation programme [7].

Data transformation also involves smoothing, aggregation, generalization, normalization, attribute construction. Data transformation routines convert the data into appropriate forms for mining.

3) Data integration:-

It combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes, or flat files. There are a number of issues to consider during data integration [4]. This is referred to as the entity identification problem. Databases and data warehouses typically have metadata - that is, data about the data. Such metadata can be used to help avoid errors in schema integration. Redundancy is another important issue. Inconsistencies in attribute or dimension naming can also cause redundancies in the resulting data set. Some redundancies can be detected by correlation analysis [4].

4) Data reduction:-

Data reduction is the process of minimizing the amount of data that needs to be stored in data storage environment. Data reduction can increase storage efficiency and reduce costs [8]. It eliminates redundant data and use data compression for reduces size of files [7]. Principle component analysis, feature selection and cluster analysis are some techniques of data reduction. Data reduction reduces the representation of data but produce the same result.

Comparison Table:-

S.no.	Publication	Year	Method	Strength	Weakness
1	Springer	2002	Gene selection for cancer classification using support vector machine	1.Used for both small and large number of data sets 2. Find the nested subsets and optimum gene.	Over fitting
2	Journal of machine learning	2003	Supervised feature selection	Improved performance	Over fitting
3	IJABIDM	2005	Rule extraction method	More accurate compare to other methods	Used only for large data sets
4	IEEE	2006	Dimensionality reduction algorithms	Feature extraction and Feature selection for large scale	Cannot use for highly complex data sets
5	IJOCS	2006	Data pre-processing for supervised leaning	1.Reduce cost 2.Feature wrappers are used	1. Slow 2.Can not give best performance for each data set
6	Springer	2008	Selective pre-processing	It can be use for both minority and majority classes.	Low Performance
7	IEEE	2012	Mega trend diffusion	1.Class possibility 2.Synthetic attributes	1. Over fitting 2.Performance

Motivation:-

Data pre-processing is the important step of data mining to resolve the problem of noisy data, inconsistent data, and redundant data. In our survey we analyses different methods and we find some problems related to pre-processing techniques. The main problems on which we are focusing are: Over fitting, performance and speed.

The motivation of this paper is how to improve data quality from which the performance of classifier can be improved.

Proposed Work:-

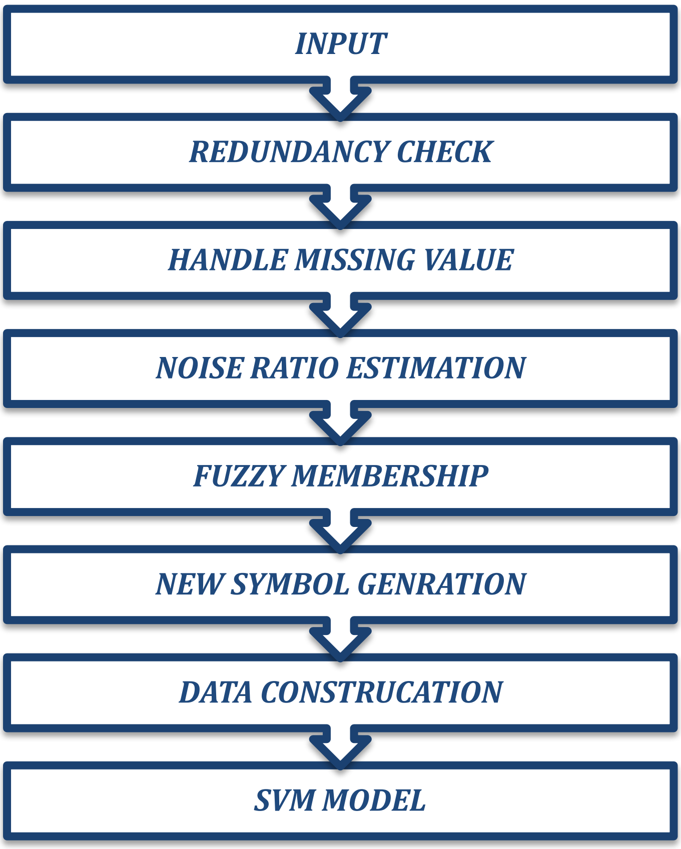

Fig.2 The proposed method

Fig. 2 is the flow chart of proposed method that starts from input data which content noise. After selection of input data we apply redundancy check then handle missing values and then we calculate noise ratio and evaluate fuzzy membership and then we generate new symbols then we construct data which is improved in quality and finally the SVM model building.

This paper deals with the small datasets whose quality we want to improve. The main elements which affect the quality of data are redundancy, missing values and noise. We apply different checks to improve the quality and reduce these problems of data.

Input: Input is a small dataset which consist redundant and noisy data. In this paper we improve the quality of this data to increase performance.

Redundancy Check: Redundancy is duplication of data in same dataset. To eliminate redundancy we analyse complete data set and remove that data which are redundant.

Handle Missing Values: Handling missing values is the important part of effective modelling. Missing values signify different things: field is not applicable, event did not happen, data is not available etc. [15]. In any dataset missing values cannot be taken so eliminate the missing values.

Noise Ratio Estimation: The main problem in dataset is noise so, to eliminate the noise we calculate noise ratio. Noise is a random error measured in variables, for example in a group discussion session if all the members start speaking together then instead of actual information it becomes noise.

We analyse dataset then noise ratio is calculated when data is overlapped. Data is in different classes and as we cannot calculate the class to which data belongs that's why to solve this problem we calculate noise ratio. While identifying noise, each instance values are calculated. Noise ratio is calculated using linear regression analysis. Linear regression analysis is an approach for modelling the relationship between a scalar dependent variable and one or more explanatory variables [7].

Fuzzy Membership: Fuzzy membership function is a curve that defines how each point in the input space is mapped to a membership value [16]. Fuzzy membership is calculated according to their participation in class or individual ratio values. After calculating fuzzy membership we generate new symbol for overlapped area. We use triangular shape membership.

New Symbol Generation: The new generated symbol can be calculated by dividing the addition of the values of data in different classes by the number of reputations at individual instance.

Data Construction: Data construction replaces the new symbols at the place of old noisy data.

SVM Model: The data is set for SVM to form the classification model. Support vector machine is a powerful classifier which uses kernel method for high accuracy.

We are going to compare both the methods proposed method and mega trend diffusion method (base paper method) to analyse which is better.

Conclusion:-

In data mining and knowledge processing three main steps are involve to find hidden knowledge from the data. First pre-processing, model building and finally testing therefore, to improve the performance of model in all three steps improvements can be adoptable. This proposed study work provides a technique by which quality of data can improve the classifiers accuracy. Which is a permissible approach for improving the performance of classifiers, that are currently used in image processing application, the proposed scheme thus also helpful for different other kind and format of data too. In future the proposed technique is implemented using MATLAB and performance of the system is provided.

References

1. R.Andonie, “Extreme Data Mining: Inference from Small Datasets,” Int. J. of Computers, Communications and Control, ISSN 1841-9836, Vol. V(2010), No. 3, pp. 280-291.
2. http://www.laits.utexas.edu
3. http://dataminingwarehousing.blogspot.in
4. http://slidewiki.org
5. http://www.tutorialspoint.com
6. www.informatik.hu-berlin.de
7. Wikipedia
8. http://searchdatabackup.techtarget.com
9. Der-Chiang Li and Chiao-Wen Liu, “Extending Feature Information for Small Data Set Classification,” IEEE Transaction on Knowledge and Data Engineering, Vol. 24, No. 3, March 2012.
10. Asa Ben-Hur, Jason Weston, “A User’s Guide to Support Vector Machines,” Department of Computer Science Colorado State University.
11. Rayner Alfred, “Optimizing Feature Construction Process for Dynamic aggregation of Relational Features,” Journal of Computer Science, 2009 Science Publication.
12. Andreas Weingessel, Martin Nattery Kurt Hornik, “Using Independent Component Analysis for Feature Extraction and Multivariate Data Projection” University of Vienna, August 1998.
13. Dustin Boswell, “Introduction to Support Vector Machines,” August 2002.
14. Bharti M. Ramageri, “Data Mining Techniques and Applications,” Indian Journal of Computer Science and Engineering.
15. https://msdn.microsoft.com
16. http://radio.feld.cvut.cz/matlab/toolbox/fuzzy/fuzzytu3.html
17. Xiujufu, lipo Wang”Data dimensionality reduction with application to improving classification performance and explaining concepts of data sets”int j. business intelligence and data mining vol 1,no .1 2005.
18. Jerzy Stefanowski1 and SzymonWilk” Selective pre-processing of imbalanced data for improving classi¯cation performance”springer 2008.
19. Jun Yan, Benyu Zhang, Ning Liu, Shuicheng Yan, Qiansheng Cheng, Weiguo Fan, QiangYang,Wensi Xi, and Zheng Chen” Effective and Efficient Dimensionality Reduction for Large-Scale and Streaming DataPreprocessing” IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 18, NO. 3, MARCH 2006.

20. S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas” Data Preprocessing for Supervised Leaning” INTERNATIONAL JOURNAL OF COMPUTER SCIENCE VOLUME 1 NUMBER 2 2006.
21. ISABELLE GUYON,JASON WESTON, STEPHEN BARNHILL,VLADIMIR VAPNIK “Gene Selection for Cancer Classificationusing Support Vector Machines” Machine Learning, 46, 389–422, 2002
22. Kluwer Academic Publishers.13] Isabelle Guyon, Andr´eElisseeff” An Introduction to Variable and Feature Selection” Journal of Machine Learning Research 3 (2003).