www.ijecs.in

International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 5 May 2017, Page No. 21289-21291
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i5.15

Novel Preprocessing Techniques for NID3R

***Trilok Suthar*[1]**

[1]M.Phil. Research scholar, C U Shah University, Surendranagar

Abstract-Stream data mining is the process of excerpting knowledge structure from large, continuous data. For stream data, various techniques are proposed for preparing the data for data mining task. In recent years stream data have become a growing area for the researcher, but there are many issues occurring in classifying these data due to erroneous and noisy data. Change of trend in the data periodically produces major challenge for data miners. Column oriented data storage has shown fast access to data used in data mining. Various missing value replacement algorithms are implemented using MonetDB. This research also concentrates on incremental missing value replacement for stream data. The proposed method generates the value for the missing data considering the data type and data distribution. It also considers the concept drift in the data stream. The method will applied to different datasets and promising results will be expected

I INTRODUCTION

As data varying over time, predictors need to have a better chance to amend or retrain themselves; otherwise they will become incorrect. The most of the predictors estimate that data came already preprocessed or data preprocessing is an integral part of a learning algorithm. If there is inappropriate and repetitious information presents in data, then data mining during the training aspect is more crucial. Real data are usually incomplete, absence of attributes, noisy and holds outlier thus it needs to be preprocessed the data. Preprocessing of data is used to advance the algorithm accuracy, completeness, interpretation, value added, consistency, good accessibility and timeliness. It is the process of converting data into clear, more productive, and in agreement as user needs. More skillful outcome and less computation time can be used as indicators. The data also becomes shorter without changing the information in it. Data preprocessing can take an extensive amount of processing time. The result of data preprocessing is the final training set. It mainly involves missing value replacement, transformation, normalization and discretization. Many supervised learning approaches that adapt to changes in data distribution over time is also called concept drift so, there are many issues created for incremental data preprocessing such as high dimensionality, online streaming, size of data and storage of these data. As Data is emerges timely, learning models need to be able to adapt the changes automatically. This research presents incremental approach of different data preprocessing techniques for stream data. As part of preprocessing techniques, missing value replacement techniques for the numeric data and categorical data are proposed. The efficiency of the proposed techniques will tested using classification accuracy. Outperforming results of the proposed methods are expected for stream data analysis.

II LITERATURE REVIEW

Data preprocessing has a great significance in data mining. Data preprocessing makes data more suitable for data mining and improve the data mining analysis with respect to time, cost and quality. Data preprocessing is most required in data mining because the data in the real world are incomplete, noisy and inconsistent. Data mining most parts are data cleaning, data reduction, data integration, transformation, discretization and normalization. Data cleaning includes missing value replacement, find the outliers and remove the noisy data. Data reduction mainly involves feature selection and heuristic method. Data integration integrates data from multiple sources such as flat files, data cubes and multiple databases. The completion of data preprocessing is the final training set [1]. The data with missing values could lead to degraded resultant accuracy. The simple way is to replace the missing value by mean value in case of numeric Literature review & analysis 4 attributes and highest frequency in the case of categorical value [2]. Sally McClean et al. invented a technique to replace the missing value by making rules establish on background knowledge but still lose some usable rules [3]. Jau Ji Shen et al. favored Rule Recycle bin technique which rehash and compose the rules to receive further outright attributes value association rule which empower the database reborn to prior the veracity and integration rate and progress the validity of missing value completion [4]. Thomas et al. proposed that an existing fuzzy rule induction algorithm can consolidate missing values in the training method in a very common way without any need for artificial replacement of the missing values themselves [5]. Mei Ling Shyu et al. designed a framework named F-DCS for replacing missing value which obtains the basic concept of conditional probability approach. This framework can manage both nominal and numeric values with a high degree of certainty when it is distinguished with other techniques such as using minimum, average and maximum value [6]. Olga et al. implemented three methods named a Singular Value Decomposition (SVD) based method, weighted K-nearest neighbours (K-NN) and row average. K-NN and SVD based methods provide quick and proper ways of measure missing values for microarray data, though K-NN is better than SVD [7]. The missing value in the dataset can influence the performance of the classification process and it became difficult to extract the useful data from datasets. To solve this problem Anjana Sharma et al. presents three techniques such as lit wise deletion, K-NN imputation and mean/mode imputation. These techniques are applied to student records of the university and fill all the missing values. These resulting datasets are tested on C5.5 algorithm and by comparing classification accuracy proved that K-NN is better than other two [8]. R. Malaryizhi et al. recruit K-NN classifier performs superior than K-means clustering in missing value imputation. [9]. Phimmarin Keerin proposed a new methodology CKNN (cluster based K-NN) to impute missing values in microarray data [10]. The new algorithm, CKNN imputation is an extension of k nearest neighbour with local data clustering being integrated for enhances efficiency and proved that the CKNN give better results compare to normal K-NN impute method. Nirmala Devi et al. forecast the replacement of the missing value by mean and median of clusters and achieve Literature review & analysis

II PROPOSED TECHNIQUES

In this research work for missing value replacement based on Skewness sensitive technique is used. In this technique first we have to find the mean and median of overall all data.

- If median<mean then it is a right skewed and all missing value will be replace by the average of all values which is less than mean.
- If median>mean then it is left skewed and all missing value will be replace by the average of all values which is greater than mean.

Proposed Skewness sensitive technique

Algorithm :SST

```
1. Input Statement
2. If data type is continuous
3. If missing value
3.1 find the incremental mean and incremental median.
3.2 If median<mean, then it is a right skewed and range is extended in the lower
side of the median. All missing values will be replaced by the average of all
the values less than mean
else
3.3 If median>mean, then it is left skewed and range is extended in the higher
side of the median. All missing value will be replaced by the average of all
the values greater than mean & repeat step1.
4. Update the corresponding latest maximum, latest minimum and incremental
standard values in Mastertable
5. else find the interval
6. Update correspondence count in dqmatrix
7. Update the output table
8. repeat step 1
9. else
8. (for categorical data)
10. If missing value
11. Find the maximum count from dqmatrix else go to step 13
12. Replace the missing values with the
correspondence categorical value of
maximum count
13. else Update correspondence count in dqmatrix
13. Update the output table
14. Repeat step 1
```

IV NID3R ALOGRITHM

NID3R is a modified ID3 [37], tree based classification algorithm, which uses CAIR/CAIM criterion for an attribute selection instead of information gain in ID3 and information gain ratio in C4.5 algorithm. Main idea is to reduce the computational complexity and improve the classification accuracy. Information gain or gain ration is calculated by considering all the class values while CAIM takes only the maximum count, so it reduces the computational complexity and gives better result than information gain or gain ratio. So in the NID3algorithm CAIM is used as online discretization during tree based classification model preparation. CAIM gives better relationship between class and attribute than information gain CAIR is used for the attribute selection stage on model preparation. Following algorithm explains the process:

Algorithm:

1. Select an attribute except that attribute whose value has to be predicted.
2. If attribute is continuous, discretize it using CAIM.
3. Calculate CAIR/CAIM value for that attribute. (Equation (1) and (7))
4. Repeat steps 1 and 2 for each attribute.
5. Then select an attribute for which CAIR/CAIM is maximum
6. Make node containing that attribute.
7. Then on the basis of that attribute, divide the given training set in to subsets.
8. Then recursively apply the algorithm on each subset until the set contains instances of the same class. If the set contains instances of the same class, then return that class.

V RESULTS

Table 1 & 2 describes the results of the missing value replacement techniques on Department & Iris database respectively.

Table 1: Classification accuracy of NID3 algorithm on Department Information Database

Algorithm Name	Minimum	Maximum	Average	Skew-ness Sensitive

				Techniqu e
NID3R	40.00%	46.66%	46.00%	46.00%
NID3R with Increment al phase	73.33%	73.33%	80.00%	80.00%

Table 2: Classification accuracy of NID3 algorithm on Iris Database

Algorithm Name	**Minimu m**	**Maximu m**	**Averag e**	**Skew-ness Sensitive Techniqu e**
NID3R	61.33%	62.66%	56.66%	62.66%
NID3R with Increment al phase	61.33%	62.66%	56.66%	62.66%

VI CONCLUSION

The incremental missing value replacement techniques are proposed for different data types and the effect of the techniques are tested using the classification accuracy of NID3R algorithm. The improved classification accuracy in most of the cases shows the superiority of the proposed techniques.

VII REFERENCES

[1]http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining- 3.html

[2]Cw.flek.cvut.cz/lib/exe/fetch.php/cources/ac4m33sad/2_tutorial.pdf.

[3] S. McClean, B. Scotney and M. Shapcott, "Using Background Knowledge with Attribute- Oriented Data Mining", Knowledge Discovery and Data mining (Digest no, 1998/310), IEE colloquiumon, pp. 1/1-1/4, 1998.

[4] J. Shena and M. Chen, "A Recycle Technique of Association Rule for Missing Value Completion" in Proc. AINA'03, pp. 526-529, 2003.

[5] Thomas R. Gabriel and Michael R. Berthold, "Missing Values in Fuzzy Rule Induction", Systems, Man and Cybernetics, IEEE International Conference on (Volume: 2), 2005.

[6] M. Shyu, I. P. Appuhamilage, S. Chen and L. Chang, "Handling Missing Values via Decomposition of the Conditioned Set", IEEE Systems, Man, and cybernetics society, pp. 199-204, 2005.

[7] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman, "Missing value estimation methods for DNA microarrays", Bioinformatics 17 (6): 520-525, 2001.

[8] Anjana Sharma, Naina Mehta, Iti Sharma, " Reasoning with Missing Values in Multi Attribute Datasets" ,International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 5, May 2013 .

[9] R. Malarvizhi, A. Thanamani," K-NN Classifier Performs Better Than K-Means Clustering in Missing Value Imputation", IOSR Journal of Computer Engineering (IOSRJCE), vol. 6, pp. 12-15, Nov. - Dec 2012.

[10] Phimmarin Keerin and Werasak Kurutach, Tossapon Boongoen, "Cluster-based KNN Missing Value Imputation for DNA Microarray Data", IEEE International Conference on Systems, Man, and Cybernetics COEX, Seoul, Korea, October 14-17, 2012.

AUTHOR PROFILE

Trilok suthar received the B.E degree in Information science and engineering from visvesvarya technological university belgavi in 2012 and m.tech degree in computer science and engineering in 2014.