Incremental Missing Value Replacement Techniques for Stream Data

Kinnari Patel
M.Tech at Manipal
University Jaipur, India

R G Mehta
Associate Prof. at SVNIT
Surat, India

M M Raghuvanshi,
PhD, Principal at RGCET,
Nagpur, India

N N Vadnere
Research & Quality
Analyst
MAYCOMS Software
Surat India

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. This research 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 is applied to different datasets and promising results derived.

Keywords

Skewness, Mean, Median, Standard deviation, Discretization.

1. 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 skilful 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 evolving over time, learning models need to be able to adapt to 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 are tested using classification algorithm. Outperforming results of the proposed methods proves their superiority for stream data analysis.

2. RELATED WORK

Data preprocessing has a great significance in data mining. It makes data more suitable for data mining and improves its analysis with respect to time, cost and quality. Data preprocessing plays major role in data mining. Data in the real world are incomplete, noisy and inconsistent so, it’s necessary to preprocess the data. Data cleaning, reduction, integration, transformation, discretization and normalization are major preprocessing tasks [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 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 skilful outcome for their classification analysis of the dissimilar number of clusters [11]. T. R. Sivapriya et al. invented new method called Least Square Support Vector Machine (LSSVM) technique based on multiple imputation pair with Z-Score and support vector machine classifier, hence proved that LSSVM is most suitable technique by comparing classification accuracy of K-NN, BPN, C4.5 and SVM [12].In real time applications, it is required to gather flow data for system over a broad range of time. It is very difficult and time consuming task to generate a data over time within the known range. To solve above problem P Gupta et al. analysed two methodologies and applied to water quantity data. The two directional exponential smoothing (TES) is used to predict the missing value for raw stream flow data and exponentially weighted moving average (EWMA) is used for seasonal data. The results show that both TES and EWMA method are giving good result and predicted the data within and outside the range of water quantity data but doesn’t work efficiently on other incremental datasets [13].

Transformation is also a beneficial step in data preprocessing. A transformation mainly implicates smoothing, generalization, aggregation; Transformation essentially restricts the maximum data. Ariel Schlamm et al. proposed data transformation technique named the Nearest Neighbour Transformation (NNT). The NNT is introduced as a data driven, preprocessing step for spectral image analysis. This approach transforms the data into a Euclidean distance space which compresses points close together and exaggerates the difference between spectrally distant points and shows good results in Clustering [14].

The missing value replacement techniques proposed so far are for static data. Stream data needs incremental strategy for deriving the new value for the missing value. The symbol transformation may come across with a new symbol, creates a need for the incremental transformation.

Different techniques for the same are proposed in the next section.

3. PROPOSED METHOD

The Stream data is a very large in size. The existing algorithms are not suitable to perform a preprocessing task in the case of concept drift. Missing value replacement methods are proposed considering the concept drift. For the smooth propcessing, the categorical values are transferred in to corresponding numeric symbols. In stream data, periodically arrival of new symbols needs incremental transformation. The proposed process performs incremental transformation and missing value replacement. For intermediate storage, master table is maintaned.

A. *preparing mastertable*

Master table is prepared to store statistical information like latest minimum value, latest maximum value, incremental meanvalue [15] and incremental standard deviation value.

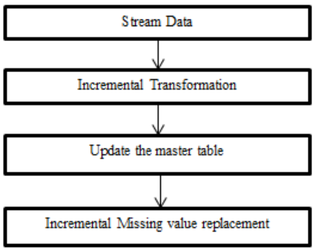
Stream Data
Incremental Transformation
Update the master table
Incremental Missing value replacement

Fig. 1 Incremental data preprocessing

B. *Incremental missing value replacement*

In this paper, different missing value replacement techniques are proposed for different data types as shown in figure 2. The missing value for the numeric field are replaced by the incremental minimum, incremental maximum or incremental mean as derived in equation (1). Missing values for the categorical data is derived using proposed Skewness sensitive technique or replaced with the attribute value with the highest latest frequency count.

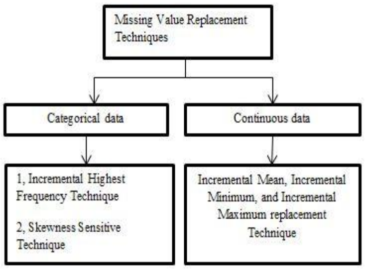
Missing Value Replacement Techniques
Categorical data
Continuous data
1, Incremental Highest Frequency Technique
2, Skewness Sensitive Technique
Incremental Mean, Incremental Minimum, and Incremental Maximum replacement Technique

Fig. 2 Incremental Missing Value Replacement Technique

a. *Incremental highest frequency technique for categorical value*

In this technique if the data are in categorical form than missing value will be replaced by the incremental categorical method that is the category which come incrementally maximum the missing value replace by that value.

b. *Incremental mean, max and min replacement technique for continuous data*

If data are continuous than it replace the missing value by incremental mean (1).

$$\begin{aligned}\mu_n &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ &= \frac{1}{n}(X_n + \sum_{i=1}^{n-1} X_i) \\ &= \frac{1}{n}(X_n + (n-1)\mu_{n-1} \\ &= \mu_{n-1} + \tfrac{1}{n}(X_n - \mu_{n-1}) \qquad (1)\end{aligned}$$

The incremental max and incremental min are latest maximum and minimum values stored in the master table, used in place of missing value. The effect of different techniques on classification accuracy is analysed in the following section.

c. *Skewness sensitive technique*

Skewness is a measure of the degree of asymmetry of a distribution. Skewness values can be positive or negative. In perfect normal distribution the tails on both sides of curve are exact mirror images of each other, in this situation mean and median are almost same. When the right tail of the curve longer than the left side, it is right skewed and all maximum frequency are concentrated on the left side. In this situation median is less than mean. Also called as the positive Skewness and when the left side of curve is longer than the right side ,it is left skewed means Skewness is more on the left side and it is left skewed and all maximum frequency are concentrated on the right side. In this situation median is greater than the mean. Also called as the negative Skewness.

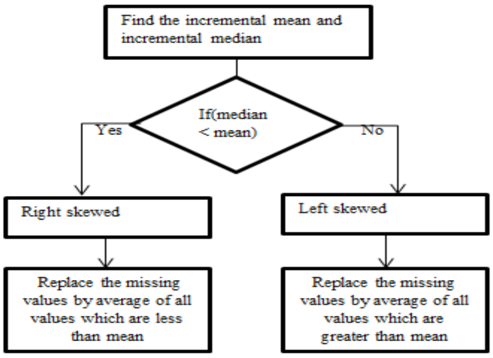
Find the incremental mean and incremental median
If(median < mean)
Yes
No
Right skewed
Left skewed
Replace the missing values by average of all values which are less than mean
Replace the missing values by average of all values which are greater than mean

Fig. 3 Skewness sensitive technique

Skewness sensitive technique finds the incremental mean from equation 1, and then finds the incremental median from the count of dqmatrix of CAIM discretization [16].

- 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.
- 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.

Suppose the data is 1, 1, 1, 2, 3, 4, 4, - ("-"indicate a missing value).

Mean=2.28

Median=2

Here, median<mean, It is right skewed.

Skewness Sensitive technique will generate 1.25 for the missing value and highest frequency technique will generate 1 for the missing value.

4. SIMULATION &RESULTS

A Java net bean is used to perform the proposed methods. Data sets are obtained from UCI repository [17] and Weka tool [18].

The proposed missing value replacement techniques are tested on various datasets, where missing values are generated randomly. Table 1 contains list of data sets for categorical data, table 2 contains datasets with continuous data. Table 3 contains list of datasets with categorical data which are generated from continuous data using CAIM discretization. These datasets are tested on NID3 classification algorithm which uses CAIR as attribute selection criterion [19].

Table 1. Dataset Description

Dataset	Feature	Instance	Missing Values	Training data
Vote	16	435	1123	50
Contact lances	4	24	30	5

Table 2. Dataset Description

Data Set	Feature	Instance	Missing Values	Training data
Housing	13	506	2887	100
Diabetes	8	768	1248	100

Table 3. Dataset Description

Data set	Feature	Instance	Missing values	Training data
Sea data	3	2500	598	500
Hyper	11	10000	7773	1000
Horse colic	25	300	1509	100
Credit rating	15	690	1583	100

B. Effect of missing value replacement technique on classification accuracy

The data sets listed in table 1 were prepared for classification using Incremental and non-incremental highest frequency techniques. The measured classification accuracy for the data sets are listed in table 4. The missing values replaced with latest highest frequency value stored in the master table are showing superior result compared to the relative static information. As from Table 4 it is seen that for Vote data and Contact lance data, the classification accuracy is improved.

Table 4. Difference between Highest Frequency Technique and Incremental Highest Frequency Technique.

Dataset	Highest Frequency Technique	Incremental Highest Frequency Technique
Contact lance	41.66%	66.66%
Vote	63.21%	65.97%

Table 5 shows the classification accuracy of NID3 algorithm for the data sets prepared using static minimum , static maximum and static mean from training data, static mean of the training data and value generated using incremental mean derived in Equation 1. For non-linearly distributed data, incremental value will be most suitable for the missing value.

Table 5. Difference between Static and Incremental Technique

Dataset	Diabetes	Housing
Incremental Mean	63.54%	84.76%
Mean	62.76%	60.07%
Incremental Maximum	56.77%	75.88%
Maximum	56.77%	75.88%
Incremental Minimum	64.06%	83.00%
Minimum	64.06%	82.60%

A. Comparison of proposed Skewness sensitive method and incremental highest frequency method.

Proposed Skewness based missing value replacement technique was applied to a sea data described in table 3. The resultant dataset is applied to NID3 algorithm. Same way the missing values are replaced using incremental highest frequency technique and the resultant data is applied to NID3 algorithm. The accuracy is measured after a block of 500 instances. The Resultant classification accuracy is depicted in the table 6. The superior results of proposed Skewness sensitive technique is shown in table 6. The Fig 4 indicates the consistent superior result for the proposed technique.

Table 6: Difference between Highest Frequency technique and Skewness sensitive technique.

Records	Highest Frequency method	Skewness sensitive Technique
500	72%	72%
1000	67.7%	70.2%
1500	66.26%	70.66%
2000	66.7%	70.10%
2500	67.88%	69.92%

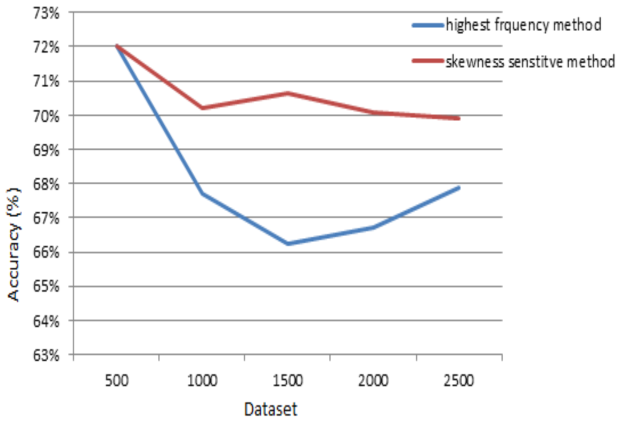

Fig. 4: Difference between Highest Frequency technique and Skewness Sensitive technique.

Proposed method is also applied on other datasets listed in table 7.

Table7. Difference between Highest Frequency Technique and Skewness Sensitive Technique.

Dataset	Highest Frequency method	Skewness Sensitive Technique
Diabetes	71.87%	71.87%
Housing	87.15%	87.15%
Hyper	50.07%	64.03%

Because of symmetrical data distribution in Diabetes and Housing dataset, getting same result with both method.

Classification algorithm is applied on two different dataset. First on the original dataset and the accuracy is measured. Secondly in the same dataset missing values has been created and proposed method is applied on that than the result has been tested. The resultant accuracy is almost same for both datasets shown in Table 8. The results show the efficiency of the missing value replacement techniques on different dataset.

Table 8. Difference between Original dataset and dataset after applying Skewness Sensitive Method

Dataset	Original data	Skewness sensitive
House	87.15%	86.95%
Diabetes	71.87%	71.87%
Horse colic	82.66%	81.66%
Credit rating	44.49%	44.49%

Other methods are very much sensitive for data type and data distribution and need to do statistical analysis in that but proposed Skewness Sensitive technique is data insensitive and there is no need to do statistical analysis, as the method performs it inherently, Results gives optimum solution.

5. CONCLUSION

Stream data needs incremental preprocesssing techniques like transformation and missing value replacement. The proposed framework makes use of incremental transformation for categorical symbols from string value to the numeric one. 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 NID3 algorithm. The improved classification accuracy in most of the cases shows the superiority of the proposed techniques.

6. REFERENCES

[1] Data preprocessing. CCSU 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, 1998, pp. 1/1-1/4.

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

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

[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 (2001) 17 (6): 520-525.

[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”, 2012 IEEE International Conference on Systems, Man, and Cybernetics October 14-17, 2012, COEX, Seoul, Korea.

[11] N. Devi, Balamurugan.S, Swathi U.V, “An amalgam KNN to predict Diabetes Mellitus”, in proc. ICE-CCN, 2013, pp. 691-695.

[12] T.R.Sivapriy, A.R.Nadira Banu Kamal, V.Thavavel, “Imputation and Classification Of Missing Data Using Least Square Support Vector Machines – A New Approach in Dementia Diagnosis”, International Journal of Advanced Research in Artificial Intelligence, Vol. 1, No. 4, 2012.

[13] Prakash Gupta, R.Srinivasan,”Missing Data Prediction and Forecasting for Water Quantity Data”, 011 International Conference on Modeling, Simulation and Control “IPCSIT vol.10 (2011) © (2011) IACSIT Press, Singapore.

[14] Ariel Schlamm, David Messinger, “Improved detection and clustering of hyper spectral image data by preprocessing with a Euclidean distance transformation”, *WHISPERS 2011, IEEE, Lisbon, Portugal, June* (2011)

[15] Tony Finch, “Incremental calculation of weighted mean and variance”, February 2009.

[16] Lukasz A. Kurgan,Member, IEEE, and Krzysztof J. Cios, ”CAIM Discretization Algorithm”, Senior Member, IEEE, IEEE transactions on knowledge and data engineering, vol. 16, no. 2, february 2004

[17] UCI repository dataset, “http://archive.ics.uci.edu/ml/”

[18] Weka tool “http://www.cs.waikato.ac.nz/ml/weka/Vote data”, “Contact lance data.

[19] M.R.Lad, R.G.Mehta, D.P.Rana, “A novel tree based classification”, international journal of engineering science & advanced technology”, ijesat | may-jun 2012.