2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing

Pre-Processing Methods of Data Mining

Asma S aleem
Department of Computer Science and Engineering
University of Engineering & Technology
Lahore, Pakistan
asmasaleem85@yahoo.com

Khadim Hussain Asif
Department of Computer Science & Engineering
University of Engineering and Technology
Lahore, Pakistan
asifkhad@yahoo.com

Ahmad Ali
Department of Bioscience
COMSATS Institute of Information Technology
Sahiwal, Pakistan
ahmadali@ciitsahiwal.edu.pk

Shahid Mahmood Awan
Al-Khawarizmi Institute of Computer Science
University of Engineering & Technology
Lahore, Pakistan
shahidawan@kics.edu.pk

Mohammed A. AlGhamdi
Institute of Innovation and Entrepreneurship
Umm Al-Qura University
Makkah, Saudi Arabia
maeghamdi@uqu.edu.sa

***Abstract*—Data generation, handling and its processing have emerged as the most reliable source of understanding and discovery of new facts, knowledge and products in the world of natural and material sciences. The emergence of the most efficient techniques in statistical or bioinformatics situations has therefore become a routine practice in research and industrial sectors. Under practical conditions, dealing with large datasets, it's likely to have inconsistencies and anomalies of all kinds to prevent to know real outcomes for practical problems. For accurate data mining computer based techniques of data pre-processing offer solutions that help the data under processing to conform normal structures which in turn considerably improve the performance of machine learning algorithms. In this process, accurate determination of outliers, extreme values and filling up gaps poses formidable challenges. Multiple methodologies have therefore been developed to detect these deviated or inconsistent values called outliers. Different data pre-processing techniques discussed in this paper could offer most suitable solutions for handling missing values and outliers in all kinds of large datasets such as electric load and weather datasets.**

Keywords—data pre-processing; data mining; outliers; missing values.

I. INTRODUCTION

The success of machine learning (ML) for a particular task is significantly influenced by the quality of given data and its demonstration. Any defective, noisy, superfluous and inappropriate data may lead to defective results [1]. In machine learning problems the data sanitation including its preparation and filtration may take up to 80% of the time on the data pre-processing in any real world project of data mining. Hence pre-processing of all data sets remains a fundamental factor for data mining problems [2], [3].

Real time data gathered through instruments, sensors or processes etc. are likely to pick up anomalies or measuring errors due to instrumental failure, problems of linking to other systems while delivering the results. The collection of incorrect and noisy data would normally result from human or computer errors at the time of entry. Additionally the data transmission, technology limitations such as limitation on buffer size during data synchronization, transfer, and consumption also perform a significant part in the quality of data. Similarly the conflicts in naming conventions or formats of data input attributes and unavailability of inputs at the time of data recording may also result in incorrect data collection.

In order to sanitize the data from such inconsistencies and anomalies different techniques of data pre-processing play a major role [2]. Knowledge discovery in database (KDD) would therefore essentially contain data pre-processing step that would help to lower data complexity and to enhance its accuracy while analysing the data [4]. This paper proposes new techniques for data pre-processing particularly for electric load and weather data to achieve following objectives:

1) Incorporation of missing values in datasets to smooth (weather or load) curve.
2) Efficient detection of outliers in datasets to reduce its level of noise.

978-1-4799-7881-6/14 $31.00 © 2014 IEEE
DOI

In the end, comparison of different combinations of data pre-processing methods is presented.

II. Literature Review

The process of data mining is an integral part of developing more understanding about different disciplines of interest for today's scientists. For this purpose data banks have been established in the form of large cyber repositories, databases and warehouses. However the major issues with this process is the extended data preparation and processing time due to inconsistencies, redundancies of the data at the source which in turn would influence the performance of data mining algorithms anywhere.

A. Management of the missing values

The most common issue that a researcher may come across in the process of knowledge development (KDD) from databases through data mining includes the missing values. A dataset with 1 – 5% missing values may not significantly influence any outcomes whereby from 5 – 15% range would need to apply sensitive algorithms/techniques to manage the quality of outcomes. Any scale of missing data beyond 15% is likely to influence the outcomes significantly [5].

The absence of single or multiple values of variable under consideration is commonly attributed as missing value [6]. Missing data, being non-persistent in nature, hence considered to be one of the most important statistical and design problems in research [7]. This is particularly true with the *mega and giga data* situations such as genomics, microarray or weather origins.

Different methodologies and techniques has been proposed from time to time to overcome missing data problems starting from very simple *List Wise Deletion* [8] to a situation where the data miners need to apply more sophisticated techniques to minimize the ill effects of data gaps. These may include *Pairwise Deletion* [9], [9], *Single-Value Imputation, Mean Mode Imputation (MMI)* [8], *Hot and Cold Deck Imputation (HDI, CDI)* [5], [8] *and K-Nearest Neighbour (KNN)* [11]. These methodologies may be more effective under one situation whereas lesser in another data mining situation. The KNN technique is referred to be the most efficient when applied on microarray data situations [12].

B. Management of Outliers

Outliers may be those observations in a dataset that do not follow any pattern or distribution and extend way outside the normal limits of a particular observation set [13], [14]. These values in the data may arise from a range of physical measurements, data transmission and quality reasons. Some of the factors may also originate from natural variations, outside contaminations or intentional manipulations. Normally in case of abundant presence of such values in the dataset, it is advisable to repeat the process or may need extra studies.

Before the preparation of data for analyses, the detection of outliers can ensure and enhance the quality of outcomes and its subsequent interpretation. Some of the important techniques used for this purpose may include *Z-Score Method* [15]*, Modified Z-Score Method* [16] with a difference that Modified *Z-Score Method* offers more robust and accurate identification of outliers especially for small datasets. A Z value > 2.5 – 3.0 would be an outlier. The *Box Plot* [17] *and 2-Sigma method* use Inter Quartile Range (IQR) and "2 Sigma" Z-score for efficient identification of outlier values to prevent these values to influence the quality of outcomes.

III. Problem Statement

As discussed in preceding text that the raw data collected from different sources is inherently vulnerable to have anomalies such as gaps, noise and other inconsistencies. The challenge is to minimize first of all, the sources of data anomalies. For example in weather or load data discussed in this paper, the most common factors include meter malfunctioning, network failure, equipment failure, maintenance and human issues.

These problems may produce significant deviation from actual data. In this paper, we applied pre-processing techniques (filling of missing values and outlier detection) on load and weather data.

IV. Proposed Methods To Handle Missing Values

The electric load data was collected from National Transmission & Despatch Company (NTDC) Limited, Pakistan and the weather data was gathered from online website [18]. Many factors like electric load consumption, time factor, weather data and possible customer classes happen to influence the electric load consumption.

Missing data is one of the most important concerns in such type of forecasting. The information contained in the attributes with missing data values is very important in the process of data analysis. The learning process on each instance is necessary as it may contain some exceptional information. Missing data normally arises in almost all serious statistical analyses. A number of approaches are available to deal with missing data situations as discussed earlier. These approaches were applied to our electric load and weather data to repair it for missing values.

A. Missing in Electric Load data

Main attributes of electric load dataset include year, month, day, hour of the day and demand for electricity. This data was dealt with the missing values in demand attribute using available algorithms. The dynamics of electricity demand depends upon the type and time of the day with electricity consumption peak would normally be in summer months and hot days compared to winter days. And for time, generally observed trend would be the day time rather than morning and evenings. A demand curve would normally show that peak demand arises in the afternoon during

summer months. Similarly the type of the day such as working day, weekends, holidays, or events are the other major factors influencing the load. Hence the data of previous and next week is deemed to be very important while finding missing values for any week day.

1) Hourly missing data

This may be a randomly missing data.

Algorithm 1:

h-1 = data of previous hour
h+1 = data of next hour
h = Average (h-1, h+1)

2) Whole Day Missing Data

The missing electric demand data of whole day may be treated by calculating the average of previous week and next week data of same day-hour.

Algorithm 2 (Type based):

for (i: 1 to 24)
h-1 = same hour of same day from last week
h+1 = same hour of same day from next week
h_i = Average (h-1, h+1)
if (i > 24)
exit; else continue;

B. Missing Weather data

Data collected from online source was based on hourly data of a day. In this data we observed outliers as well as missing values. Firstly, outliers were identified by previously discussed methods then missing values handled by following algorithm.

The nature of missing data in the weather may be similar as the load data; however methods to find these missing values in case of weather data may be different. The missing values in weather data are also concerned with the missing of whole data row. In this case scenario, filling the gaps for missing data rows would be the first step before moving further to impute missing values in these data rows.

It is a common observation that weather condition of current day is mostly dependent on the consecutive days or hours rather than the type of the day (as in the case of electric load data), hence slightly different approaches are used to find the missing values in weather data.

1) Data of Whole day is Missing:

If the weather data for whole day is missing then the following given algorithm is used form filling missing values.

Algorithm 3:

To treat this type of missing values in weather dataset, the data from previous and next day is used to find the average.

h-1 = data of previous day
h+1 = data of next day
h = Average (h-1, h+1)

2) Consecutive Hours Missing Data

It is a bit tricky to deal with missing data of consecutive hours and preferably weights are used to find the missing values. The weights are assigned to both values that are being used to compute missing value. The weights are assigned in such a way that the nearest value has the greatest impact than the other value in calculating missing values.

Algorithm 4 (Weighted Average):

Let Position of missing value = p
Total No. of values used to compute missing value = N
Weight for 0 value = w2 = 1/N
Weight for no.1 value = w1 = 1- w2
Missing value at no. p = M.V = (w1 * value at no. p-1) + (w2 * value at position N-1)

Example:

Following gives an example to calculate missing value using weighted Average (WA) algorithm.

TABLE I. EXAMPLE OF WEIGHTED AVERAGE (WA)

No.	*Values*	*Missing Values*
1.	*2*	
2.	*0*	*2.25*
3.	*0*	*2.4984*
4.	*0*	*2.7924*
5.	*3*	

Missing value at no.2 = 2*(.75) + 3*(.25) = 2.25
Missing value at no.3 = 2.25*(.666) + 3*(.333)
= 1.4985 + .9999 = 2.4984
Missing value at no.4 = 2.4984*(.50) + 3*(.50)
= 1.2492 + 1.5 = 2.7492

V. OUTLIER DETECTION METHODS

The outliers or unpredicted data points can also provide valuable information. Ozone hole data was first recorded as outliers however, the management of such data points sometimes becomes difficult to whether keep or remove it and set it down as experimental error.

Methods used to identify outliers have been reviewed in the preceding sections. These missing values are then filled by the Weighted Average (WA) and Type Based (TB) algorithms that are proposed in this paper.

VI. RESULTS AND ANALYSIS

Different combinations of methods for outlier detection and missing values were applied to both weather and load datasets. The results obtained were then compared to give the best combination of pre-processing methods to yield smooth curve of load and weather data.

A. Combinations of Weighted Average (WA) Algorithm with Outlier Detection Methods

Following curve was drawn from the raw (unprocessed) dataset of weather. The data of weather used for this paper

was from 1st April, 2014 to 16th April, 2014 of Lahore city of Pakistan from online weather source [18]. Only *'Temperature'* attribute was considered for applying these pre-processing techniques. It contained almost 695 number of observations.

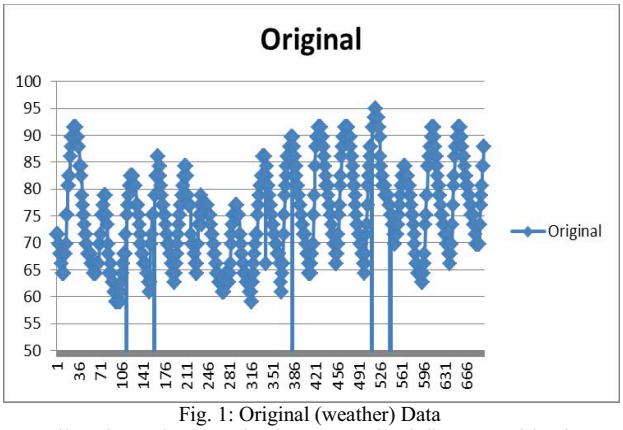

Fig. 1: Original (weather) Data

Outlier detection method was applied first, on this dataset to recognize the outliers followed by the application of Weighted Average (WA) algorithm to detected outliers.

1) Z-Score and Weighted Average (Z-WA)

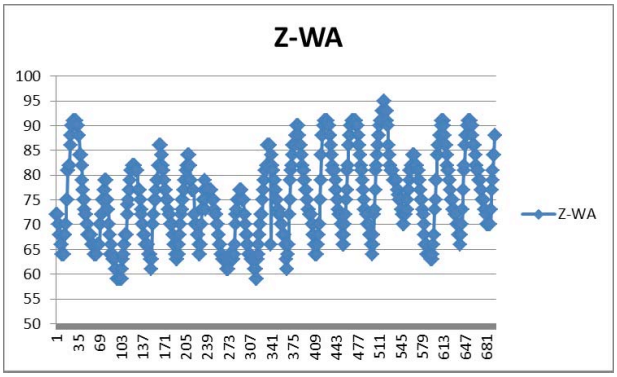

Fig. 2: Combination of Z-Score And Weighted Average (Z-WA)

2) Modified Z-Score and Weighted Average (ModZ-WA)

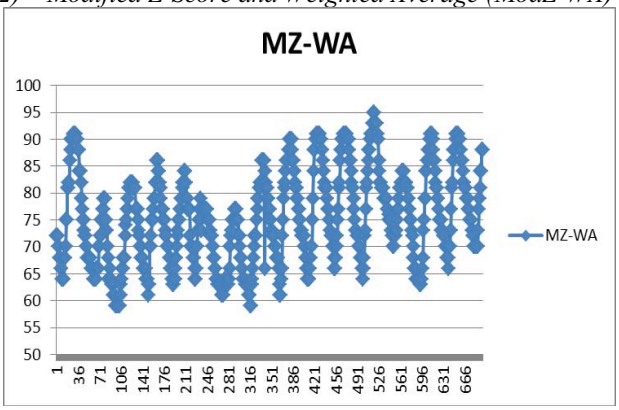

Fig. 3: Combination of Modified Z-Score And Weighted Average (ModZ-WA)

3) Box-Plot and Weighted Average (BP-WA)

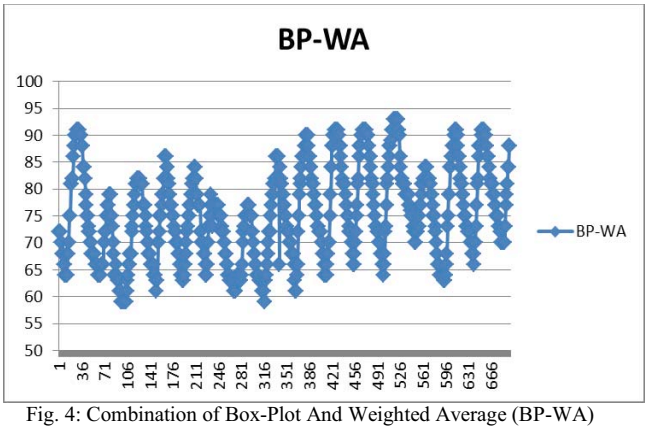

Fig. 4: Combination of Box-Plot And Weighted Average (BP-WA)

4) 2-Sigma and Weighted Average (2S-WA)

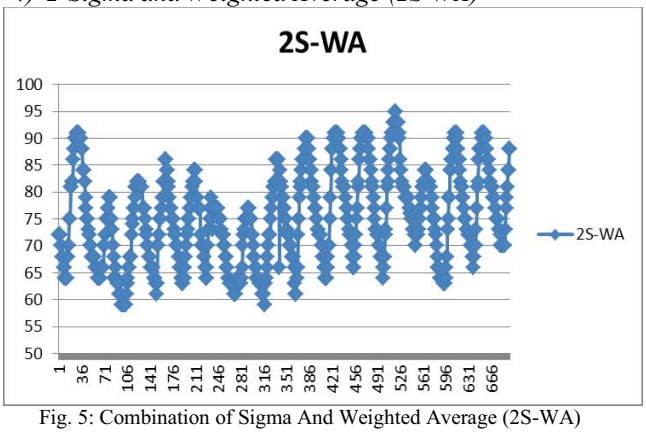

Fig. 5: Combination of Sigma And Weighted Average (2S-WA)

B. Combinations of Type Based (TB) Algorithm with Outlier Detection Methods

Following curve resulted from the raw (unprocessed) dataset of load collected from NTDC Limited, Pakistan. There were about 17612 observations of load data under consideration.

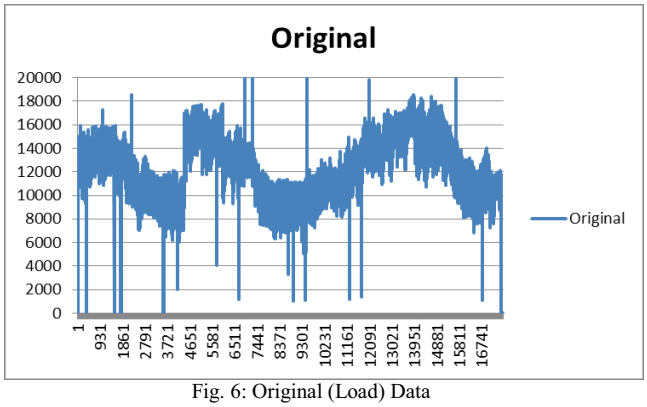

Fig. 6: Original (Load) Data

1) Z-Score and Type Based (Z-TB)

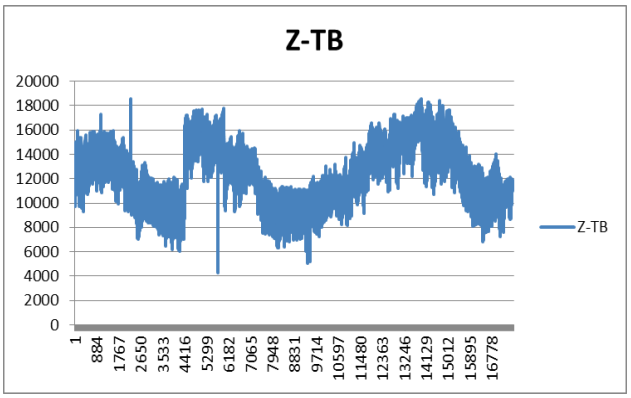

Fig. 7: Combination of Z-Score And Type Based (Z-TB)

2) Modified Z-Score and Type Based (ModZ- TB)

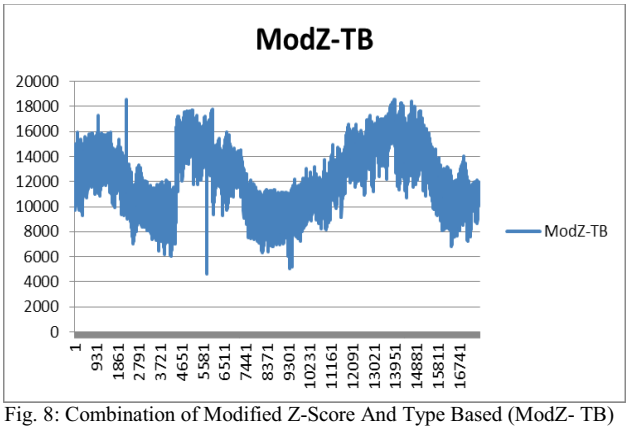

Fig. 8: Combination of Modified Z-Score And Type Based (ModZ- TB)

3) Box-Plot and Type Based (BP- TB)

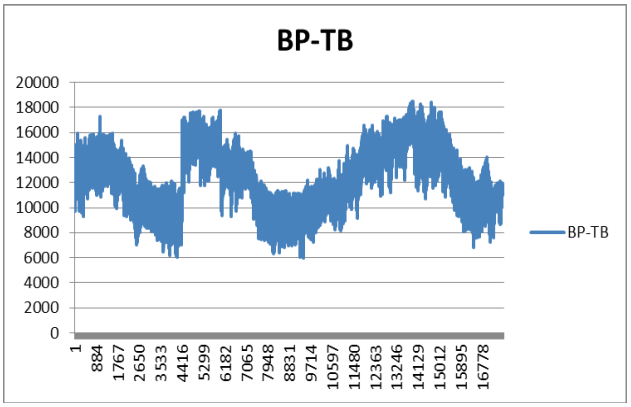

Fig. 9: Combination of Box-Plot And Type Based (BP- TB)

4) 2-Sigma and Type Based (2S- TB)

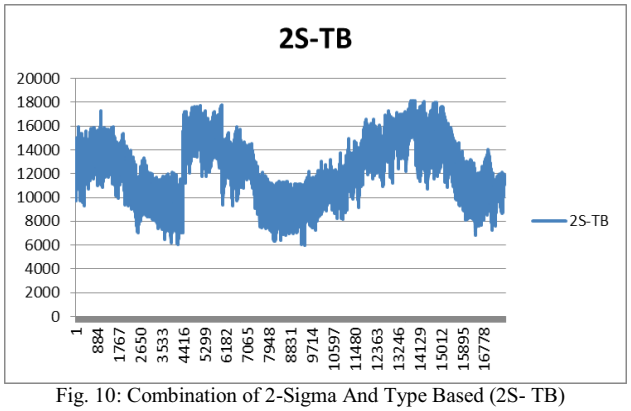

Fig. 10: Combination of 2-Sigma And Type Based (2S- TB)

C. Comparison of Results

The results obtained from different combinations were compared to the original curve. The following section demonstrates the results from combinations of outlier detection methods with missing value techniques to identify the best combination.

1) Comparison of Original (Weather) Data with Combinations

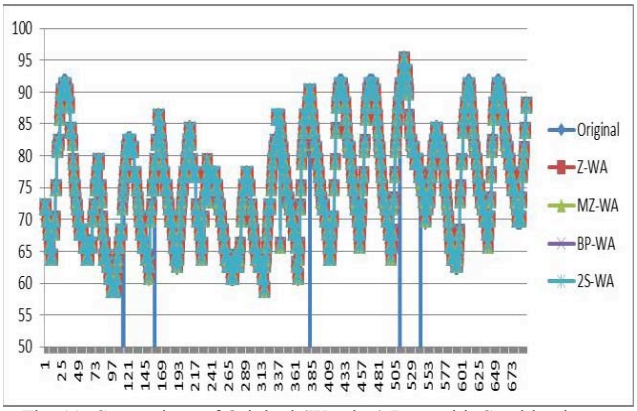

Fig. 11: Comparison of Original (Weather) Data with Combinations

2) Comparison of Original (Load) Data with Combinations

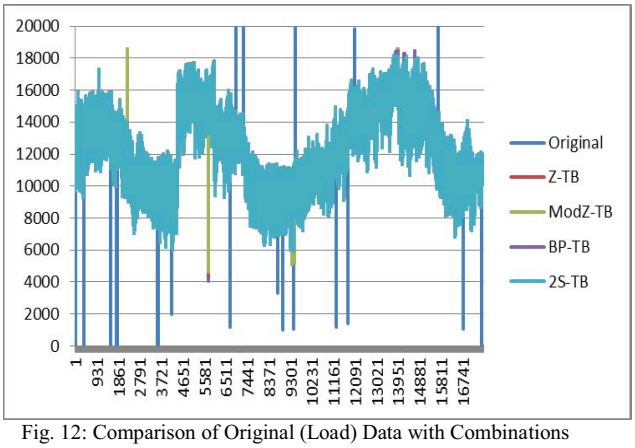

Fig. 12: Comparison of Original (Load) Data with Combinations

D. Analysis of Results

This section gives the facts and figures for the analysis of results and comparisons given in the previous section.

1) Analysis of Weighted Average (WA) Algorithm with Outlier Detection Methods

TABLE II. ANALYSIS OF WEIGHTED AVERAGE (WA) ALGORITHM WITH OUTLIER DETECTION METHODS

No. of Observations		695
Outlier Method	**No. of Detected Outliers**	**No. of Missing Values**
Z-Score	5	5
Mod Z-Score	5	5
Box Plot	7	7
2-Sigma	5	5

2) Analysis of Type Based (TB) Algorithm with Outlier Detection Methods

TABLE III. ANALYSIS OF TYPE BASED (TB) ALGORITHM WITH OUTLIER DETECTION METHODS

No. of Observations		17612
Outlier Method	**No. of Detected Outliers**	**No. of Missing Values**
Z-Score	141	141
Mod Z-Score	142	142
Box Plot	151	151
2-Sigma	164	164

VII. CONCLUSION & FUTURE WORK

A. Conclusions:

From the comparison of graphs produced in section VI, it can be concluded that combinations of Weighted Average (WA) and Type Based (TB) algorithms with box-plot (BP) and 2-Sigma (2S) methods gives the best results to smooth the curve for further processing.

For load dataset, box-plot (BP) and 2-Sigma (2S) methods are deemed to best for the detection of outliers while Type Based (TB) algorithm outperforms for filling missing values in the case when the data of whole day or consecutive hours is missing.

For weather dataset, box-plot (BP) and 2-Sigma (2S) methods proved to be best methods to effectively detect the outliers while Weighted Average (WA) algorithm outperforms for filling missing values. Weighted Average (WA) algorithm can also be applied for filling missing values in load datasets when missing data occurs randomly.

B. Future Work:

The data pre-processing techniques that we have proposed in this research paper are application specific, as these are developed for filling missing values in load and weather datasets for load forecasting system. These proposed algorithms can be further extended for general purpose applications. Other applications in which these algorithms can be applied are forecasting systems. Examples of forecasting applications are weather forecasting and meter reading forecasting systems.

REFERENCES

[1] Tom M. Mitchell, "The Discipline of Machine Learning", School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, July 2006.

[2] S. B. Kotsiantis, D. Kanellopoulos and P. E. Pintelas, "Data Preprocessing for Supervised Leaning", International Journal of Computer Science, Volume 1 Number 2, 2006, ISSN 1306-4428.

[3] Mohammed J. Zaki and Wagner Meira Jr., "Data Mining and Analysis: Fundamental Concepts and Algorithms", Cambridge University Press, 2013.

[4] Joel H. Levine and Thomas B. Roos, "Introduction to Data Analysis: The Rules of Evidence", Macintosh HD: DA: DA IX: Volume I: 006 Intro (What is the wealth), March 1997.

[5] Acuna E. and Rodriguez C., "The treatment of missing values and its effect in the classifier accuracy". In D. Banks, L. House, F.R. McMorris,P. Arabie, W. Gaul (Eds).Classification, Clustering and Data Mining Applications. Springer-Verlag Berlin-Heidelberg, 639-648. 2004.

[6] Paul D.Allison, "Missing Data", December 2008.

[7] Azar, B., "Finding a solution for missing data", Monitor on Psychology, 33, 70, 2002.

[8] Liu Peng, Lei Lei, "A Review of Missing Data Treatment Methods".

[9] Judi Scheffer, "Dealing with Missing Data", Res. Lett. Inf. Math. Sci. (2002) 3, 153-160.

[10] John Graham, "Missing Data: Analysis and Design", Chapter 2, 2012.

[11] K. Hron, M. Templ, and P. Filzmoser, "Imputation of missing values for compositional data using classical and robust methods", December, 2008.

[12] Danh V. Nguyen, Naisyin Wang and Raymond J. Carroll, "Evaluation of Missing Value Estimation for Microarray Data", Journal of Data Science 2(2004), 347-370.

[13] D.M.Hawkins. "Identification of Outliers". Chapman and Hall, 1980.

[14] Chen, Z. Fu, A. & Tang, J., "Detection of Outliered Patterns", Dept. of CSE, Chinese University of Hong Kong, 2002.

[15] Schiffler RE. "Maximum Z Score and outliers", The American Statistician, Vol. 42, No.1(Feb., 1988), 79-80 .

[16] Iglewicz, B., Hoaglin, D. "How to detect and handle outliers", ASQC Quality Press, 1993.

[17] Turkey, JW., "Exploratory data analysis", Addison-Wesely, 1977.

[18] Weather Data Source: (http://www.wunderground.com/)