ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

KNN with Pruning Algorithm for Simultaneous Classification and Missing Value Handling

S.Visalakshi, V.Radha

Research Scholar, Dept of Computer Science, Avinashilingam Institute for Home Scienceand Higher Education for Women, Coimbatore , India.

Professor, Dept of Computer Science, Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

Abstract - In real-time dataset, lots of missing values are prompt and habitually become a very serious problem in data mining. A dataset with missing value becomes a general problem of data quality. Some sort of missing data are present moreover in training set or testing set which will affect the accuracy of any learned classifiers. Handling of missing data is very significant, as they have a pessimistic blow on the interpretation and the result of data mining process. Handling of missing value techniques can be grouped into a bunch of four categories namely, Imputation methods, Maximum Likelihood, Machine Learning and Complete Case Analysis. The K-Nearest Neighbor (KNN) is one of the imputation techniques used to treat missing value. The drawback of KNN is overcome in the proposed KNN. The proposed technique is implemented and used for identify contamination in drinking water. Missing value is implemented in the water dataset. It helps to improve automatic water contamination detection scheme. Some of the existing and popular imputation techniques are compared, and it is proved that the proposed system produces better results compared to other imputation techniques.

Keywords - Potable water, Missing value handling, Imputation, Hot deck, Mean, K-Nearest Neighbor, Contamination detection, Data Pruning.

I. INTRODUCTION

The process of mining hidden knowledge from database is called data mining. Data mining is capable of extracting hidden patterns from the dataset, predicting the future trends and knowledge driven decision. Water is one of the natural resources used by all living and non-living beings. Potable drinking water is safe and sound enough to be consumed by humans. The potable water supplied to households, commerce and industry meet drinking water standards [1]. In many parts of the world, it is difficult to access potable water and the public are still using contaminated water for daily routine. As the use of non-potable water for food preparation leads to widespread acute and chronic illnesses, it has become a major cause of death and misery in many countries. In developing countries, reduction of waterborne disease is a major goal. Contamination can be intentional and non- intentional [2]. The main goal of the research is to identify the contamination present in drinking water.

Real-life data is surrounded by noise, is inconsistent and is incorrect. Many existing industrial and research data contains missing values. Data preparation, cleaning and transformation comprise the majority of work in all kinds of data mining applications. The major task in data pre-processing is Data Cleaning (filling in missing values, smoothing noisy data, identifying or removing outlier and resolving variation), Data Integration (integrating the database from multiple sources (n-number of databases, data dice or documents), Data Transformation (process normalization and aggregation), Data Reduction (reduce the volume but no change in analytical results) and Data Discretization (reduction in part of the dataset especially for numerical datasets). The reason for missing value includes manual data entry, equipment errors and incorrect measurements. In real time, missing value handling is categorized into four groups, namely, Complete Case Analysis, Imputation Methods, Maximum Likelihood Method and Machine

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Learning Methods. The most common problem faced in data mining and database is missing values. If the missing value rate is <1% then it can be trivial, 1-5% is manageable, 5-15% need some sophisticated methods to handle missing values and if the missing value rate is > 15% then the dataset must be handled carefully with sophisticated tools [3]. Missing values may generate bias and affect the quality and performance of the classification algorithm. In this paper, experimental and research work concentrates on handling missing values of the real time data which are collected from distribution point. Sample water is collected from different distribution points for detecting contamination and the data contains many missing values. This paper focuses on handling missing value and removing outlier (anomalies) from the data. The experimental result of this paper gives information on how the missing value and outlier detection is performed. Imputation means replacing a new value wherever the data is missing in the database. [12]. According to Pedro J.et.al, the KNN imputation method search for more similar instances in the entire dataset. Hence, it takes more time for finding similar instance. (To find new value for the replacement of missing value).To overcome this problem, clustering is introduced to search a neighbor value for replacement, and it is illustrated with an example of real-time data. The performance of the proposed algorithm has been compared with existing imputation techniques and it performs very well. Also the result shows improvement. In Section II the literature study is discussed and in Section III the proposed techniques are explained. Section IV reports the experimental results of the enhanced KNN imputation technique. Finally, Section V presents the conclusion and scope for future study.

II. LITERATURE STUDY

The first task of data mining is pre-process of the data. Missing attribute values becomes more common problem in real world data sets. The most important task in data pre-processing is to find the missing attribute values. Missing values occur during data collection, repeated diagnosis test, experimental set and so on. Several studies have been focused on developing algorithms that deal with missing data [4, 5, 6].Generally, missing values in data mining are handled using the following three different ways:

- Discard the missing attribute value,
- Maximum likelihood approaches, and
- Imputations of missing attribute value.

Jerzy W. Grzymala-Busse and Witold J. Grzymala-Busse analyzed the two different methods of handling missing values. They are sequential and parallel methods [7]. Insequential method, missing value is replaced with known value, but in the parallel method knowledge is retrieved directly without imputing the missing values from the original dataset. Sequential methods include deleting cases with missing attribute values, replacing missing attribute value with the most common value of attributes, assigning all possible values for missing attribute, replacing the mean value of the numerical attribute, closest fit case, etc. In parallel method, missing value will be considered during the process of acquiring knowledge. The method includes Learning from Examples Module, Version 2 (LEM2), C4.5 ("don't care") and CART. Also, the author presents the work based upon the percentage of missing values. The handling of missing value techniques varies from one dataset to another dataset. To handle the missing attribute values, the best possible methods or universal methods should be preferred, using the criterion of optimality and multi-fold cross validation experiments.

Yoshikazu Fujikawa and TuBao Ho, analyze the existing methods to deal with missing attribute values and the report is given along with evaluation of processing cost and quality of imputing missing values. Three new cluster based methods are also proposed to impute missing values and the result shows that proposed cluster-based algorithm works effectively [8].

Hai Wang and Shouhong Wang propose a framework for rough set rule generation method which will enable the user to mine patterns to attain an association rule. The author also proposes four different approaches to deal with missing data [9]. The first approach for missing value is to eliminate all the missing value from the dataset. The second approach is to estimate the missing value in the dataset (Imputation). The third approach is to use common "unknown attribute value" for missing data. The fourth approach is treating the missing data as non-deterministic data. Pattern of missing data is also discussed briefly, through associated rule induction. A general pattern of missing data includes set-off, closing; hiding and disclosing are expressed by prototypes of association rules. Rough sets rule induction technique is a powerful tool for discovering knowledge.

Fourteen different imputation approaches for classification are presented [10]. The following three different problems arise with missing values in data mining. They are loss in effectiveness; obstacle in handling and investigate the data; unfairness resulting

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

from differences between missing and complete data. If the dataset is having any incomplete data either in the training set or in testing set and might be present in both (training and testing), it is sure that prediction accuracy of the classifier is degraded. Three sections of classification are used to categorize them and examine the best imputation strategies. The literature study for classification with various imputation methods is investigated and comparison is made for different imputation methods. This concludes that imputation method helps to fill the missing value better than the case deletion method.

Bhavisha.et.al explains about imputation methods like Parametric, Non-Parametric and Semi-Parametric methods [11].

A novel KNN imputation using feature weighted distance, based on mutual information, was proposed by [13], which helps to improve the classification performance of missing data. It surveys the popular imputation methods such as Mean and Mode Imputation (Mimpute), HotDeck Imputation (HDImpute) , KNN and Prediction model.

III. TECHNIQUES FOR MISSING VALUE

The KNN algorithm is one of the popular approaches to handle missing data problem. KNN imputation (used to estimate missing value for imputation) employs the k-nearest neighbor algorithm to estimate and replace missing data**.** The main advantages of KNN are that it is capable of estimating both the qualitative and quantitative attributes. It is not possible to create a predictive model for each attribute with missing data. To perform KNN, it needs three different variable values: making a decision on how many neighbors are taken for estimation in each of the iteration, training data and metric to measure the closeness. The main factor of this method is distance metric. The main problem in KNN is that, it will check the whole dataset to find similar instance for imputation. Selecting the "k" value and the measure of similar instance will reflect the result greatly. The above problem is solved by introducing clustering (pruning algorithm) before it starts searching for the k-nearest neighbor value.

The proposed method for missing value is presented in Figure 2 below. The data is given as input to the system. To remove the outlier, pruning algorithm (Clustering) is applied. Once the data is pruned, classify the dataset into complete and incomplete instances. The main purpose of introducing pruning algorithm is to reduce the time complexity, and thus automatically speed up the process of handling missing values and classification.

The procedure for pruning algorithm is given below:

1) First step includes: Identify all clusters and divide into small and large clusters with the help of K-Mean clustering. The small clusters are removed from the dataset.
2) The second step works only with large clusters and uses relevancy metric called Cluster Validity Index (CVI) to identify irrelevant data. The CVI is good performance measure if the diversity of the subsets is high. When large portion of features are irrelevant, then it is likely to get a couple of clusters with similar information separate by the unimportant features. The parameter mainly used to identify weak clusters that do not have any influence on the final result.
3) Most important part of the proposed pruning algorithm is the selection of irrelevant cluster. This is achieved by applying K-means algorithm iteratively with different parameters to create subset of clusters. The CVI is then used to identify weak clusters with irrelevant data and merge similar clusters together. The Quality of cluster is evaluated by using Cluster Validity Index which is shown in following Equation 1.

$$CVI = E_k \left\{ \frac{\sum_{i,j \in \Delta_k} D_{i,j}}{\sum_{i,j \in \Delta_k, j \notin \Delta_k} D_{i,j}} \right\} \quad (1)$$

where Ek(.) represents expectation over k and is a set showing subsets in the k^{th} cluster. In the clustering context, the CVI shows whether there are significant clusters in the data or not.

4) When clusters are completely separated, CVI is very small, near zero. In the worse case, when there are no distinct clusters, CVI is a value that is near to one. While combining with missing value imputation, this index shows whether a specific collection of subsets represents a complex or simple missing value pattern. Simple missing value patterns correspond to well-separated subset clusters and consequently smaller CVI.
5) A clear separation is generated by using 1/CVI on a linear scale when CVI is small. The more samples per feature are found by the clustering algorithm, the better the generalization of the classifiers would be. Thus, there is a direct relationship between the quality of the clustering results and the ultimate

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

performance of the classification process. The process is repeated until the CVI of clusters remains unchanged.

In the above algorithm, the CVI measure is used to select only those clusters that have relevant features. The threshold of convergence is set to 0.01 and the number of clusters k is set to 10. The clusters obtained for each distance metric is considered as a separate subset. The algorithm can considerably reduce the time complexity and in the worst scenario, all clusters can be combined after m-1 iterations. Although the pruning step does not always have a large impact on the performance, it may reduce the computational complexity.

The overall process of proposed algorithm is shown in Figure 2. Pruning method is applied to reduce the time complexity of the proposed algorithm, hence to speed up the process of missing value handling and classification. Find the nearest neighbor value using the proposed KNN in short time period. Real time water dataset which is collected from TWAD Board contains many missing values. The proposed algorithm is implemented and results are discussed in the following section.

This section concludes that missing values will be handled effectively with the help of proposed technique. Handling missing value is very important in data mining applications. If the missing values are more than 15%, then handling details difficult. To obtain accurate results, efficient techniques are needed to handle the missing values. The proposed and enhanced KNN predict and replace the missing values in a very short time.

IV. EXPERIMENTAL RESULTS

Many Imputation/Non-Imputation methods are there to deal with missing data which are developed in recent years. The basic steps for handling missing values along with imputation method are discussed Section II. In this section, experimental results about missing data are discussed. The real time dataset is collected from various distribution points. More than 5 parameters are considered to detect contaminations present in the drinking water. Water must be treated properly before it is distributed to the public.

In this paper, for removing the outlier, pruning algorithm is used. The pruned data will be given as input for further process of imputation. The dataset is classified as complete and incomplete instances. The complete instances are clustered, using K-Mean clustering. The incomplete instance are arrange, based upon the missing value percentage in ascending order, to find the nearest complete instance using distance metric. The enhanced KNN is used to impute the missing values in the dataset. The above implemented enhanced KNN is evaluated, using the three objective measures such as Accuracy, Normalized Root Mean Square Error (NRMSE) and Execution Speed.

$$\text{Accuracy} = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions} \times 100$$

$$\text{NRMSE} = \frac{\sqrt{mean\ [(y_{true} - y_{imp})^2]}}{variance(y_{true})}$$

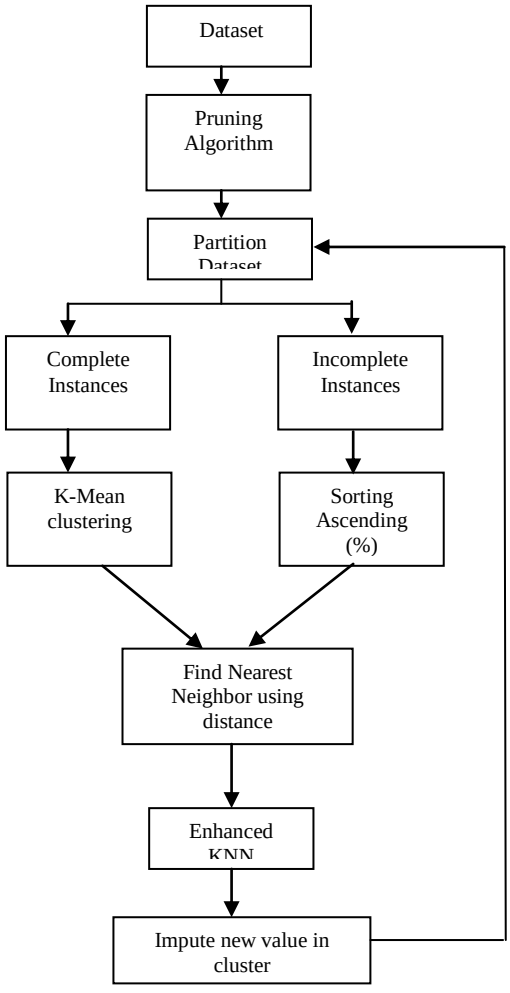

Figure.2. Proposed Missing Value Handling

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Two different datasets (summer and winter) are used for the evaluation of algorithm. The performance of the dataset is evaluated in terms of classification accuracy which is shown in Table 1. The performance of classification with high accuracy signifies the better imputation of missing values. For example, if the missing dataset is 40% during summer season, the classification accuracy of the proposed algorithm without pruning is 84.39% and with pruning it is improved to 84.97%. Similarly, for the winter season is also it is evaluated.

The Table.2. shows the evaluation of normalized root mean square error for the various seasons such as winter and summer. Generally, the lower NRMSE signifies the better imputation of missing values. For example, if the missing dataset is 40% during summer season. The error rate of the proposed algorithm without pruning is 0.4569% and with pruning is 0.4251%. Similarly for the winter season it is evaluated and shown in the same table.

Table 1. Classification Accuracy (%)

DS	MV %	M1	M2	Hot Deck	KNN	Proposed	
						WoP	WP
S	0	39.67	43.21	46.61	70.69	83.69	83.89
	10	40.24	44.27	48.12	80.78	88.67	88.92
	20	40.34	44.32	48.49	78.66	87.11	87.17
	30	40.39	44.51	48.82	75.14	86.19	86.22
	40	40.44	44.81	49.03	72.85	84.39	84.97
W	0	40.21	42.49	46.29	75.61	83.26	82.87
	10	41.41	45.44	48.55	81.23	89.46	89.51
	20	41.59	45.73	48.72	79.14	86.38	86.42
	30	41.73	45.91	48.87	77.58	85.55	85.61
	40	41.84	44.11	49.24	76.54	84.61	84.73

DS= Dataset, S=summer, W=winter, MV=Missing Value, M1=Mean, M2=Median, HD=Hot Deck, WoP=Without Pruning, WP=With Pruning.

Table 2. Normalized Root Mean Square Error

DS	MV %	M1	M2	HD	KNN	Proposed	
						WoP	WP
S	0	0.8426	0.8171	0.7881	0.5943	0.4786	0.4451
	10	0.8216	0.7981	0.7562	0.5791	0.4210	0.4049
	20	0.8296	0.7962	0.7586	0.5714	0.4318	0.4103
	30	0.8318	0.7994	0.7671	0.5521	0.4533	0.4282
	40	0.8399	0.8018	0.7699	0.5522	0.4569	0.4251
W	0	0.8566	0.8174	0.7964	0.6354	0.5519	0.4726
	10	0.8244	0.7991	0.7621	0.6033	0.4217	0.441
	20	0.8267	0.7954	0.7693	0.6109	0.4312	0.4573
	30	0.8316	0.8031	0.7769	0.6126	0.5214	0.4415
	40	0.8365	0.8078	0.7792	0.6159	0.5334	0.4548

Table.3. Execution Speed (Seconds)

DS	MV in %	M1	M2	HD	KNN	Proposed	
						Without Pruning	With Pruning
S	0	3.26	3.88	4.41	9.07	7.15	5.13
	10	3.26	3.96	4.52	9.23	7.26	5.23
	20	3.43	4.28	4.98	9.94	7.39	5.41
	30	3.66	4.91	5.47	10.21	7.82	6.28
	40	3.92	5.26	5.98	10.75	8.23	7.12
W	0	3.32	3.97	4.59	9.44	7.12	6.37
	10	3.43	4.10	4.74	9.74	7.43	6.41
	20	3.58	4.32	5.12	10.26	7.56	7.05
	30	3.71	4.97	5.73	10.79	7.94	8.26
	40	3.99	5.34	6.23	11.04	9.11	8.97

As seen in Table 3. suppose the missing dataset is 40% during summer season, the error rate of the proposed algorithm without pruning is 8.23% and with pruning is 7.12%. Similarly for the winter season is also evaluated and shown in the same table. According to the obtained results, the problem of the KNN is overcome with the help of the enhanced KNN. In this paper, the mean, median, hot deck, KNN and proposed enhanced KNN are applied to water dataset. The experimental result shows that the proposed KNN performs better for all the three metrics such as classification accuracy, normalized root mean square error and execution speed for two different seasons.

ISSN (Online) : 2319 - 8753
ISSN (Print) : 2347 - 6710

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st & 22nd March Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

V. CONCLUSION

In this paper, imputation method is used to replace the missing values. The enhanced KNN method is used to classify and impute the missing instances in the collected datasets. The proposed method selects the K-nearest neighbor, considering the input instances relevant to the target class instances. The imputation is done based on the Euclidean distance metric. During the missing value estimation, this enhanced KNN imputes missing values in a short span of time and improves the classification accuracy, NRMSE and execution speed. The result shows the best performance compared to other substitution methods. In future, the same algorithm can be used for different real time applications and implemented for database repositories.

ACKNOWLEDGEMENT

The authors express their gratitude to TWAD Board for their whole hearted support in providing dataset for research.

REFERENCES

[1] Visalakshi.S and Radha.V, "Drinking Water Quality Management and Monitoring – A Study," *International Journal of Current Research*, vol.5, issue, 10, pp.3125-3127, 2013, ISSN: 0975-833x.

[2] National Public Health Partnership and Canberra, "Framework for Management of Drinking Water Quality – application to small water supplies," *Australian Drinking Water Guidelines, Health Council*, 2004.

[3] Vinod. N.C. and Dr. Punithavalli. M., "Performance Evaluation of Mutation / Non- Mutation Based Classification With Missing Data," *International Journal on Computer Science and Engineering*, ISSN: 0975-3397, vol. 5, issue.02, pp. 56-61, 2013.

[4] Somasundaram. R.S and Nedunchezian. R, "Missing Value Imputation using Refined Mean Substitution," *IJCSI International Journal of Computer Science*, Issues, vol. 9, issue 4, No 3,ISSN (Online): 1694-0814, pp:306-313, 2012.

[5] Yoshikazu Fujikawa and TuBao Ho, "Cluster-based Algorithms for Filling Missing Values," *Japan Advanced Institute for Science and Technology, Tatsunokuchi, Ishikawa 923-1292, Japan*, 2011.

[6] Lall, U. and Sharma, A., "A nearest-neighbor bootstrap for re-sampling hydrologic time series," *Water Resource*. vol.32, pp.679–693, 1996.

[7] Jerzy W. Grzymala-Busse and Witold J. Grzymala-Busse "Handling Missing Attribute Values," *Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing, Lecturer Notes in Computer Science, Chapter 1*, vol. 3642, pp. 342-351, 2005.

[8] Yoshikazu Fujikawa and TuBao Ho, "Cluster-based Algorithms for Filling Missing values," *Asia Conference on Advances in Knowledge Discovery and Data Mining*, pp. 549-554, ISBN: 3-540-43704-5, *Springer*, 2002.

[9] Hai Wang and Shouhong Wang, "Discovering patterns of missing data in survey databases: An application of rough sets," Vol.36, Issue.3, pp.6256-6260, *Elsevier*, 2009.

[10]Julián Luengo et.al, "On the choice of the best imputation methods for missing values considering three groups of classification methods," DOI 10.1007/s10115-011-0424-2, *Springer*, 2011.

[11] Bhavisha Suthar, Hemant Patel and Ankur Goswami, "A Survey: Classification of Imputtion Methods in Data Mining," *International Journal of Emerging Technology and Advanced Engineering*, vol.2, issue.1, ISSN: 2250-2459, pp: 309-312, 2012.

[12] Gabriele B. Durrant, "Imputation Methods for Handling Item-Nonresponse in the Social Sciences: A Methodological Review," *ESRC National Center for Research Methods and Southamption Statistical Sciences Research Institure*, Universtiy of Southampton, pp. 1-42, 2005.

[13] Pedro J, Garcia-Laencina, Jose-luis Sancho-Gomex, Anibal R.Figueiras-Vidal and Michel Verleusen, "K nearest neighbours with mutual information for simultaneous classification and missing data imputation," Neuro Computing, *Elsevier*, pp. 1483–1493, 2009.