IJIRR

*International Journal of Information Research and Review*
***Vol. 03, Issue, 03, pp. 2022-2026 March, 2016***

**Research Article**

# AN ALGORITHM TO CONSTRUCT MISSING FEATURE VALUES USING K-NN ITERATIVE METHOD

**[1]*Sorna Gowri, A. and [2]Dr. Ramar, K.**

[1]The M.D.T Hindu College, Tirunelveli, Tamilnadu, India
[2]Einstein College of Engineering, Tirunelveli, Tamilnadu, India

**ARTICLE INFO**

***Article History:***

Received 08th, December 2015
Received in revised form
06th, January 2016
Accepted 13th, February 2016
Published online 31st, March 2016

***Keywords:***

WBCD, MAR,
K-NN iterative method.

**ABSTRACT**

Data mining and knowledge discovery tools become one of the foremost research areas in the field of medical diagnoses. The aim is to classify large datasets into patterns that can be used to extract useful knowledge. For example, data mining techniques can utilize patient's databases for automated medical diagnoses. The purpose is to achieve more accurate findings, speed up the diagnoses, and reduce the errors and mistakes occurred by human being. However, incomplete dataset or missing features values may affect data mining findings. The problem of missing features values is common in many applications, particularly, in medical databases. The process of treating unknown attributes values with the most appropriate values is a common concern in data mining and knowledge discovery. The process of constructing missing values is a vital process in most supervised and unsupervised data mining researches because it may affect the quality of learning and the performance of classification algorithms.

## INTRODUCTION

This paper presents a new approach for constructing missing features values based on iterative nearest neighbors and distance metrics. The proposed approach employs weighted *k*- nearest neighbor's algorithm. The main idea is to propagate the classification accuracy to a certain threshold which is set by the researchers and users. The proposed method showed slight improvement of 0.005 classification accuracy on the constructed dataset (the new dataset with no missing values) than the original dataset which contain some missing features values. The approach also showed the classification accuracy from *k*=1 to *k*=5, it showed that the maximum classification accuracy was 0.9698 when k=1.

### Related Study

The literature shows variety methods for treating the missing attribute values. These methods maybe labeled into sequential and parallel methods. In sequential methods, missing attributes values are replaced by known values then the knowledge is acquired for a dataset with all known attribute values. Examples of sequential methods are deleting the records (cases) that contain missing values, substitute missing features values with the most common value of an attribute, assigning all possible attribute values to missing features values, replacing the missing features values with the mean of feature values (Grzymala-Busse and Grzymala-Busse, 2010). Now let's have a brief attempts for treating missing features values using the above two types of methods. White did propose the simplest method to deal with missing values by simply ignoring the cases which contain unknown attributes values. Kononenko *et al.* (Kononenko *et al.*, 1984) proposed a method to observe the missing features values from other attributes. They used the class label to determine the missing attributes values. His plan was to assign the most probable attribute value $ai$ to the missing attribute $aj$ that most satisfy a class C. Another method anticipated by Quinlan (Meng and Schenker, 1999) was to employ the decision tree to estimate the missing values. The approach takes a subset $T'$ from the training set $T$. The equivalent values for the missing values in the subset $T'$ must be known. In the subset cases, the missing attribute became the class label and vice verse. Using $T'$, the decision tree can be built to determine the value of missing attributes which converted into class label (temporally). This method uses the class label to determine the missing attributes values and to utilize all the information in the case (dataset instance). However, this method is only valid when there is only one missing attribute value. Quinlan also proposed another method for handling the missing attributes values by considering the missing attributes values "*unknown*" as an actual value for the attributes. However, this solution is not valid for all the cases

***Corresponding author: Sorna Gowri, A.**
*The M.D.T Hindu College, Tirunelveli, Tamilnadu, India.*

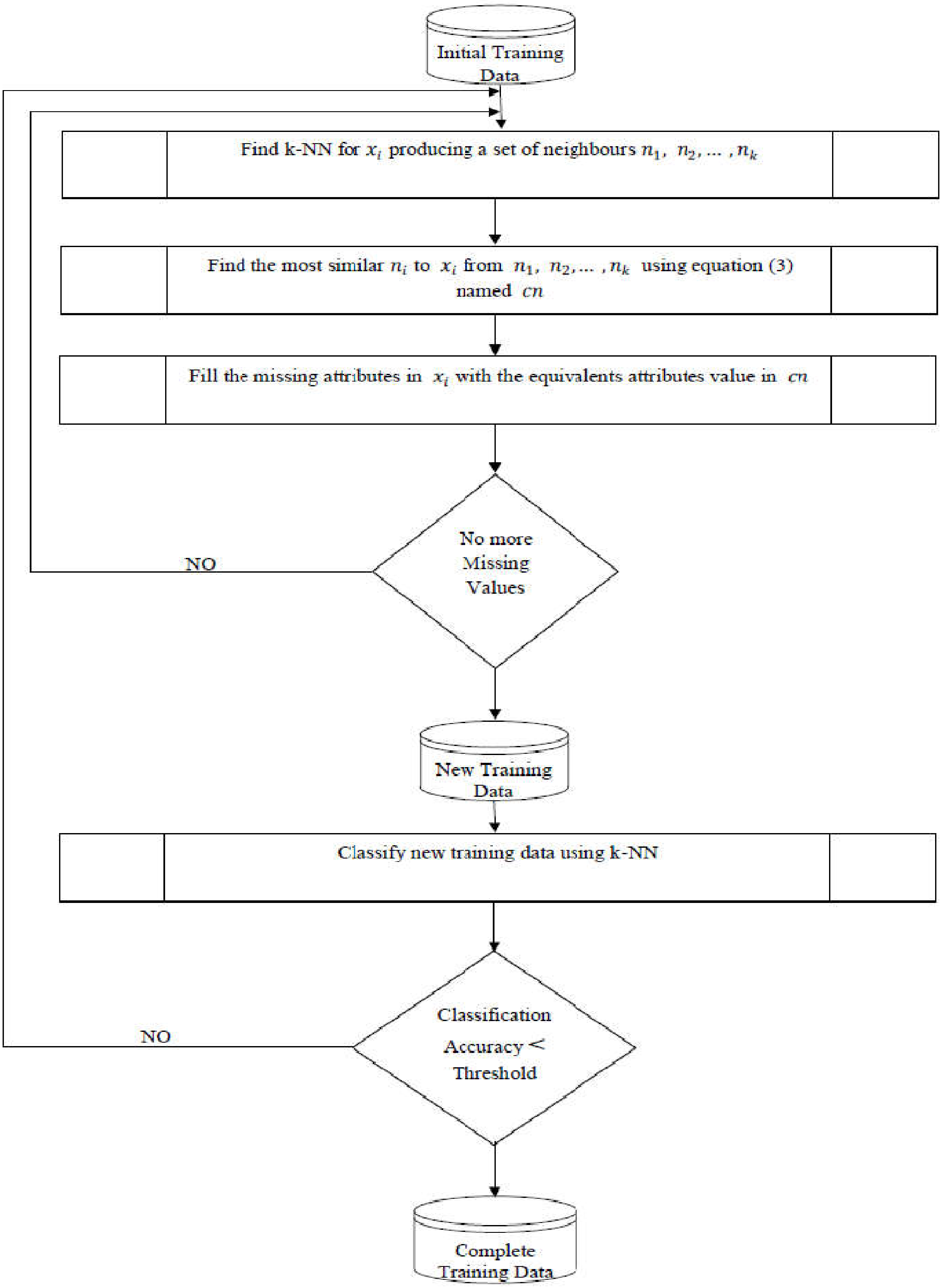
Initial Training Data
Find k-NN for $x_i$ producing a set of neighbours $n_1, n_2, \dots, n_k$
Find the most similar $n_i$ to $x_i$ from $n_1, n_2, \dots, n_k$ using equation (3) named $cn$
Fill the missing attributes in $x_i$ with the equivalents attributes value in $cn$
No more Missing Values
NO
New Training Data
Classify new training data using k-NN
Classification Accuracy < Threshold
NO
Complete Training Data

**Figure 1. The Flowchart for the proposed method (constructing missing values)**

because the value “unknown” may represent many meanings such as the value is too large or too small to be recorded, the value didn’t recorded by mistake, etc. hence, this method may bring uncertainty. Meng and Schenker (Meng and Schenker, 1999) showed that likelihood techniques can be used to deal with missing data. However, likelihood method requires specific programs which may not be available easily. Alternative to likelihood techniques is imputing the missing data.

Multiple-imputation is a method of generating multiple simulated values for each incomplete dataset, and then iteratively analyzing datasets with each simulated value substituted. The intention in this method is to generate estimates that better reflect true variability and uncertainty in the data that contains some missing values (Rubin, 2004). There are different ways to perform multiple-imputation but most approaches assume missing data to be missing at random (MAR). Missing at random (MAR) is a circumstance in which missing values are randomly distributed within one or more subsamples not among all the dataset. For example, missing more among malignant than benign but random among each class (Marshall *et al.*, 2010). Santhakumaran (Santhakumaran, 2010; Simon *et al.*, 2003) successfully used ANN to treat missing features values on WBC. The author used back propagation algorithm to train the network and used four missing value replacement methods to replace the missing values in dataset (Successive Iteration, mean, median, and mode).Among these four methods, Median method produced a promising result.

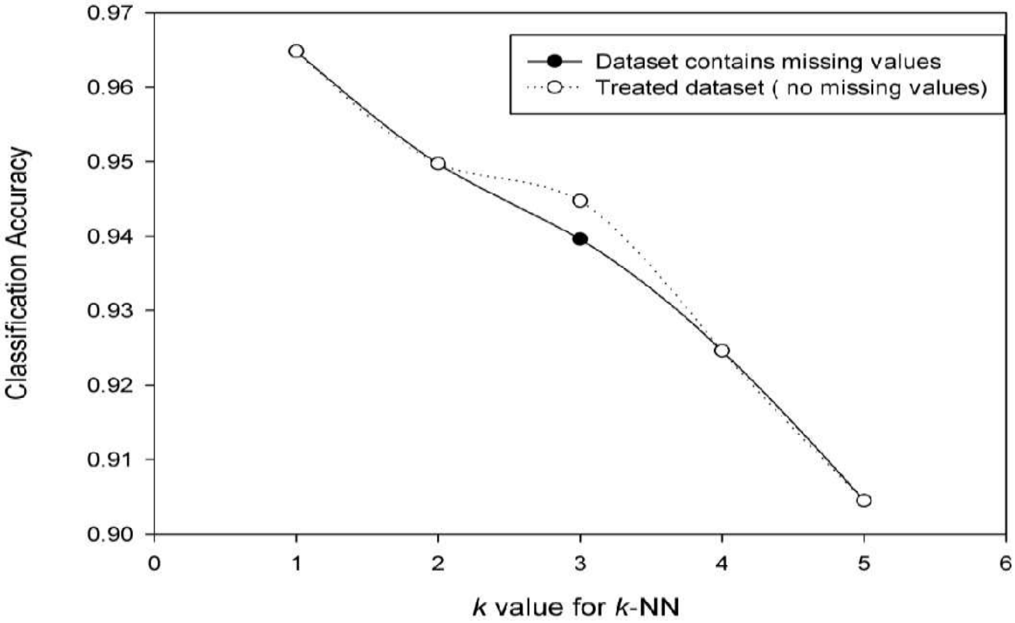

**Figure 2. A comparison of classification accuracy for the proposed method through Euclidean/*k*-NN**

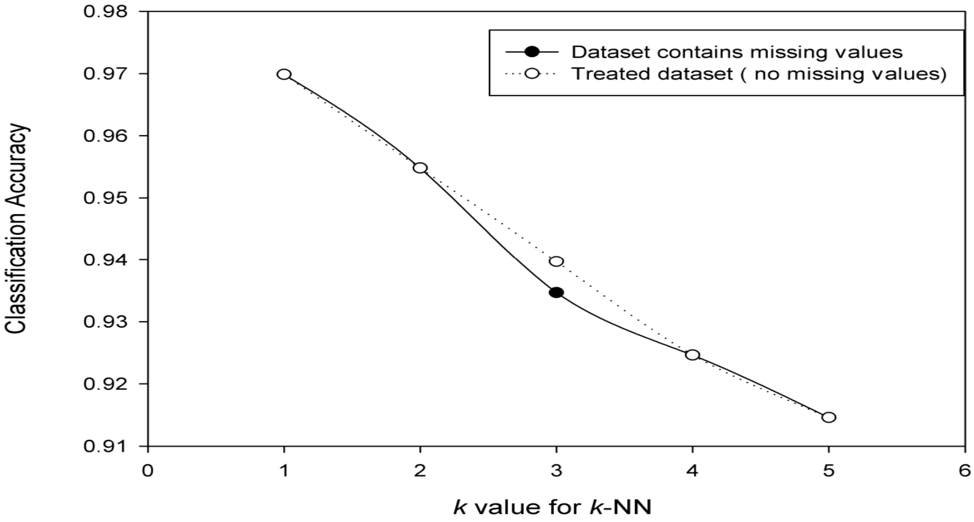

**Figure 3. A comparison of classification accuracy for the proposed method through Minkoski/*k*-NN**

## The Proposed Method

The proposed method integrates the weighted *k*-nearest neighbor's algorithm and propagating the classification accuracy to a certain threshold. The *k*-NN is to find the closest neighbors ($n1,\ldots,nk$) for a certain instance ($xi$), that contains missing feature values, using the Euclidean, Manhattan, Minkowski, and other distance functions but this work focused finally on two distance functions, Euclidean and Minkowski. Then the proposed approach is to find the most similar instance to ($xi$) from ($n1,\ldots, nk$) using formula by finding the distances values ($cni$). Where $cni$ donate to the closest neighbors to the instance $xi$, $d\ xj$, $nj$ is the distance between the instance $xj$ and the neighbor $nj$, and $nij$ is the feature $i$of the neighbor $nj$. After finding the closest neighbor (the smallest value of $cni$) call it $cn'$, the missing feature values in $xi$ will be filled by the equivalent features values in $ni$ which have $cn'$ distance to $xi$. The process of filling missing features values will produce a new training dataset (*NT*) that contains no missing features values. To verify the accurateness of the constructed missing features values, the new training dataset is applied to *k*-NN and record the accuracy. If classification accuracy is less than a threshold then the algorithm will step back to fill the missing features values until the desired classification accuracy is reached. Figure 1 shows the flowchart for the proposed method.

## The Experimental Results

In data mining and statistical researches, the split sample approach is a commonly used study design in studies that contain large dataset. This design divides the dataset into a training set and a testing set to approximate classification accuracy. The classifier is designed and developed based on the theory and then trained using the training dataset. After tainting the classifier, it is applied to each case in the testing sample. In practice, dividing data is important to avoid large bias in estimation the classifier accuracy (9). Therefore, the dataset (WBC) has been divided into two parts, training dataset and testing dataset. The dataset separation was random to avoid unfairness. The training dataset contains 500 cases 16 of them contain missing features values. The rest of the original dataset (WBC) reformed the testing cases (199 cases). After preparing the datasets, a classifier has been developed using the proposed method. The development tool was Microsoft Visual Studio 2010. The selected language was C# programming language.

In the implementation, this work has used many metrics to compute the distance including the Euclidean and Minkowski functions as a component toward the iterative *k*-NN classifier. Constructing the missing features values using the proposed method through iterative *k*-NN classifier with the Euclidean distance function showed a classification accuracy enhancement of 0.005 when *k*=3 from the first iteration and a maximum classification accuracy of 0.9648 . Figure 2 shows a comparison of classification accuracy when the missing features values were not treated and when treated. The figure also shows different classification accuracy which depend on number of neighbors (*k*) in *k*-NN. The experiment shows that more neighbors reduce the classification accuracy. The reason may due to noise as a result of more neighbors.

The experiment of constructing the missing feature values using the proposed method through *k*-NN classifier with the Minkowski distance function showed a classification accuracy enhancement of 0.005 when *k*=3 and r=1.5 from the first iteration and a maximum classification accuracy of 0.9698. Figure 3 shows a comparison of classification accuracy when the missing features values were not treated and when treated. The experiment also showed that Manhattan, Chebychev, and Canberra distance metrics are not suitable for constructing the missing attributes values, in this experiment, because the classification accuracy after treating the missing values remain lower than the classification accuracy for the original dataset.

### Conclusions

The Proposed new approach for constructing missing features values based on iterative *k* nearest neighbors and the distance functions. The approach is an iterative approach until finding the most suitable features values that satisfy classification accuracy. The proposed approach showed improvement of 0.005 of classification accuracy on the constructed dataset than the original dataset on both Euclidean and Minkowski distance functions. Manhattan, Chebychev, and Canberra distance metrics produced lower classification accuracy on the new dataset than the original dataset. This work also noticed that classification accuracy depend greatly on the number of neighbors (*k*). The experiment showed that less neighbors may lead to more accuracy. The reason for that, in my opinion, is the amount of noise produced from conflict neighbors. Finally, the maximum classification accuracy was on *k*=1 which was 0.9698.

## REFERENCES

Ashraf, M., Le, K. and Huang, X. 2010. Information Gain and Adaptive Neuro-Fuzzy Inference System for Breast Cancer Diagnoses, *in International Conference on Computer Sciences and Convergence Information Technology (ICCIT).* 2010, IEEE: Seoul. p. 911-915.

Blum, A.L. and Langley, P. 1997. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97(1–2): p. 245-271.

Buxton, B.F., Langdon, W.B. and Barrett, S.J. 2001. Data Fusion by Intelligent Classifier Combination. *Measurement and Control*, 34(8): p. 229-234.

Džeroski, S. and Ženko, B. 2004. Is combining classifiers with stacking better than selecting the best one? *Machine Learning*, 54(3): p. 255-273.

Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *The Journal of Machine Learning Research*, 3: p. 1289-1305.

Goonatilake, S. and Khebbal, S. 1994. *Intelligent Hybrid Systems*, John Wiley & Sons, Inc.

Grzymala-Busse, J.W. and Grzymala-Busse, W.J. 2010. Handling Missing Attribute Values Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, (Editors). *Springer US*. p. 33-51.

Hall, M.A. and Smith, L.A. 1997. Feature subset selection: a correlation based filter approach. 1997.

Introducing Hepatitis, C. 2012. Available from: http://www.hep.org.au/.

Kononenko, I., Bratko, I. and Roskar, E. 1984. Experiments in automatic learning of medical diagnostic rules. *in International School for the Synthesis of Expert's Knowledge Workshop*, Bled, Slovenia.

Kuncheva, L. and Whitaker, C. 2001. Feature Subsets for Classifier Combination: An Enumerative Experiment, in Multiple Classifier Systems, J. Kittler and F. Roli, (Editors). *Springer Berlin Heidelberg*. p. 228-237. 112. Grabusts, P., The Choice of Metrics for Clustering Algorithms, *in Proceedings of the 8th International Scientific and Practical Conference. Volume II*. 2011. p. 70-76. 113. Jurman, G., Riccadonna, S., Visintainer, R., & Furlanello, C., Canberra distance on ranked lists. *In Proceedings, Advances in Ranking–NIPS 09 Workshop* , 2009, p. 22-27.

Leach, M. 2012. Parallelising Feature Selection Algorithms. University of Manchester: Manchester.

Lee, S.L. 2012. Thyroid Problems. 2012 (sighted 2012 01/03/2012); Available from: http://www.emedicinehealth.com/thyroid_problems/article_em.htm.

Marshall, A., *et al*., 2010. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology*, 10(1): p. 7.

Meng, X. and Schenker, N. 1999. Maximum likelihood estimation for linear regression models with right censored outcomes and missing predictors. *Computational Statistics &amp; Data Analysis*, 29(4): p. 471-483.

Rubin, D.B. 2004. *Multiple Imputation for Nonresponse in Surveys*. New York Wiley & Sons.

Santhakumaran, F.P. 2010. An Algorithm to Reconstruct the Missing Values for Diagnosing the Breast Cancer. *Global Journal of Computer Science and Technology*, 10(2): p. 25-28.

Simon, R., *et al*., 2003. Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification. *Journal of the National Cancer Institute*, 95(1): p. 14-18.

Tsoumakas, G., Angelis, L. and Vlahavas, I. 2005. Selective fusion of heterogeneous classifiers. *Intelligent Data Analysis*, 9(6): p. 511-525.

Vijayasankari, S. and Ramar, K. 2012. Enhancing Classifier Performance Via Hybrid FeatureSelection and Numeric Class Handling- A ComparativeStudy. *International Journal of Computer Applications*, 41(17): p. 30-36

Zhang, H. and Su, J. 2008. Naïve Bayes for optimal ranking. *Journal of Experimental & Theoretical Artificial Intelligence*, 20(2): p. 79-93.

*******