LOOPING WEIGHTED K-NN ALGORITHYM FOR CONSTRUCTING MISSING FEATURE VALUES FOR CANCER DATASETS

Benard Nyangena Kiage*

Abstract

Healthcare facilities have at their disposal vast amounts of cancer patients' data. Medical practitioners require more efficient techniques to extract relevant knowledge from this data for accurate decision-making. However the challenge is how to extract and act upon it in a timely manner. If well engineered, the huge data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as cancer. Expert systems for decision support can reduce the cost, the waiting time, and liberate medical practitioners for more research, as well as reduce errors and mistakes that can be made by humans due to fatigue and tiredness. The process of utilizing health data effectively however, involves many challenges such as the problem of missing feature values, the curse of dimensionality due to a large number of attributes, and how to go about determining the features that can lead to more accurate diagnosis. Effective data mining tools can assist in early detection of diseases such as cancer. In this paper, we bring forth a new approach for constructing missing features values based on iterative nearest neighbours and distance metrics. This approach employs weighted k-nearest neighbour's algorithm. This integrates Euclidean and Minkowski functions as a component toward looping k-NN classifier. Our major inspiration is to propagate the classification accuracy to a certain threshold that is set by either researchers or users.

Key words: *Data Mining, selection, classification accuracy, neural networks, missing feature values*

* School of Computing and Information Technology; Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya

1. INTRODUCTION

Medical Databases today can range in size into hundreds of millions of terabytes. Within these masses of data lies hidden information of strategic importance. Drawing meaningful conclusions about this vast data has always been a challenge to healthcare practitioners. Data mining (DM) solves this problem. DM is a non trivial extraction of implicit, previously unknown, and imaginable useful information from data. DM finds important information hidden in large volumes of data. DM is the reasoning of data. It is the use of software techniques for finding patterns and consistency in sets of data [12]. Although computational, the utility of data mining algorithms can be used as qualitative tools to analyze quantitative data, particularly the large, complex databases being created by the health informatics community. Many countries have embraced the global healthcare system. This is done by standardizing healthcare in communication and building electronic healthcare records (EHR).

Health records may include a range of data such as general medical records, patient examinations, patient treatments, medical history, allergies, immunization status, laboratory results, Radiology images and other useful medical information for examination. This rich information may help researchers in examining and diagnosing diseases using computer techniques. EHR are capable of being shared across healthcare providers in various countries [1].

Data stored in hospital warehouses range from quantitative to analog to qualitative data; however well structured, these data conceal implicit patterns of information which cannot readily be detected by conventional analysis techniques.

Cancer diagnosing based on machine intelligence and previous history can be a step towards the reduction of the suffering of cancer patients in the entire world over. What is required however is a reliable, accurate and efficient approach for identifying diagnostic features that best describe data for the purpose of differentiating malignant and benign form of cancer, determining how missing feature values can improve prediction in determining the performance achieved by the data mining technique used and Investigating how classification accuracy and missing values can improve results by fusing the existing data mining algorithms for cancer diagnosis.

In this paper we present a new approach for constructing missing features values based on iterative nearest neighbours and distance metrics. This approach employs weighted k-nearest neighbour's algorithm. Our major inspiration is to propagate the classification accuracy to a certain threshold that is set by either researchers or users.

The structure of this paper is as follows; section 2 describes missing feature values, section 3 describes an overview of related literature, section 4 presents our method, section 5 provides the experimental results for missing feature values, section 6 conclusions and future work, section 7 provides the references used.

2. MISSING FEATURE VALUES

To achieve more accurate findings, speed up the diagnoses, and reduce the errors and mistakes that may have occurred during human intervention the missing feature values must be constructed. The construction of missing feature values is reasonably done because incomplete dataset or missing features values may affect data mining findings. The reasons for errors and mistakes of missing data or feature values includes some features values not specified because they are not available at the time of data collection, attributes value might be forgotten, mistakenly erased , or not filled during data entry.

3. RELATED LITERATURE

The related prior work from literature shows a variety of methods for treating missing attribute values. However these methods are labeled as sequential and parallel. Missing attributes in sequential methods have the values replaced by known values then the knowledge is acquired for a dataset with all known attribute values. 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 are some of these sequential methods [78].

In parallel methods, knowledge is acquired directly from the original data sets. Rule induction is an example of parallel method. Here rule learning algorithm is used to learn directly from the original dataset to find kind of regulations on how to treat or construct missing features values if they exists [78].The simplest method to deal with missing values is by simply ignoring the records containing attribute values that are unknown as proposed by White [92].

On their proposed method, Kononenko et al. [93] used the class label to determine the missing attributes values. Quinlan [94] proposed a 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 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.

Multiple-imputation [96], a method of generating multiple simulated values for each incomplete dataset, and then iteratively analyzing datasets with each simulated value was proposed to generate estimates that better reflect true variability and uncertainty in the data that contains some missing values. Santhakumaran [98] 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.

4. OUR METHOD

This study integrates the weighted *k*-nearest neighbours' algorithm and propagating the classification accuracy to a certain threshold. The *k*-NN is to find the closest neighbours ($n_1,\ldots,n_k$) for a certain instance (x_i), that contains missing feature values, using the Euclidean and Minkowski distance functions. This approach finds the most similar instance to (xi) from ($n1\ldots$) using the formula:

$$P(c_j|x_i) = \frac{P(cj).P((x_i|c_j)}{P(x_i)} \quad (1)$$

By finding the distances values (cn_i) the formula below will be applied;

$$cn_i = \frac{\sum_{j=1}^{k} {}^{n_{ij}}/_{d(x_{j,n_j})}}{\sum_{j=1}^{k} {}^{1}/_{d(x_{j,n_j})}} \quad (2)$$

Where cni denote to the closest neighbours to the instance x_i, (x_j,n_j) is the distance between the instance x_j and the neighbour n_j, and n_{ij} is the feature i of the neighbour n_j. After finding the closest neighbour (the smallest value of cn_i call it cn', the missing feature values in $x_i$will be

filled by the equivalent features values in n_i which have cn' distance to x_i. The process of filling missing features values will produce a new training dataset 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 the accuracy is recorded. 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.

5. THE EXPERIMENTAL RESULTS FOR MISSING FEATURE VALUES

In our study, the split sample approach is used because of a large dataset involved. At the same time it is important to avoid large bias in estimation of the classifier accuracy. To This end, the dataset was divided into two namely; a training set and a testing set. The purpose for these divisions is primarily to approximate classification accuracy. The design and development of the classifier is based on the theory brought forward which then is trained using the training dataset. After tainting the classifier, it is applied to each case in the testing sample. Therefore, the dataset (WBC) has been divided into two parts, training dataset and testing dataset. To avoid unfairness the dataset separation was random. The training dataset contain 500 cases where 16 of them contain missing features values. In the implementation, this work has used Euclidean and Minkowski metrics to compute the distance functions. This is 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 will show a classification accuracy enhanced. Varying k between k=1 to k=3 the iteration will show maximum classification accuracy when k=3. Linear graph tabulation will be used to compare the classification accuracy when the missing features values are not treated and when treated. This tabulation will also present various classification accuracy dependent on the number of neighbours, (*k*) in *k*-NN.

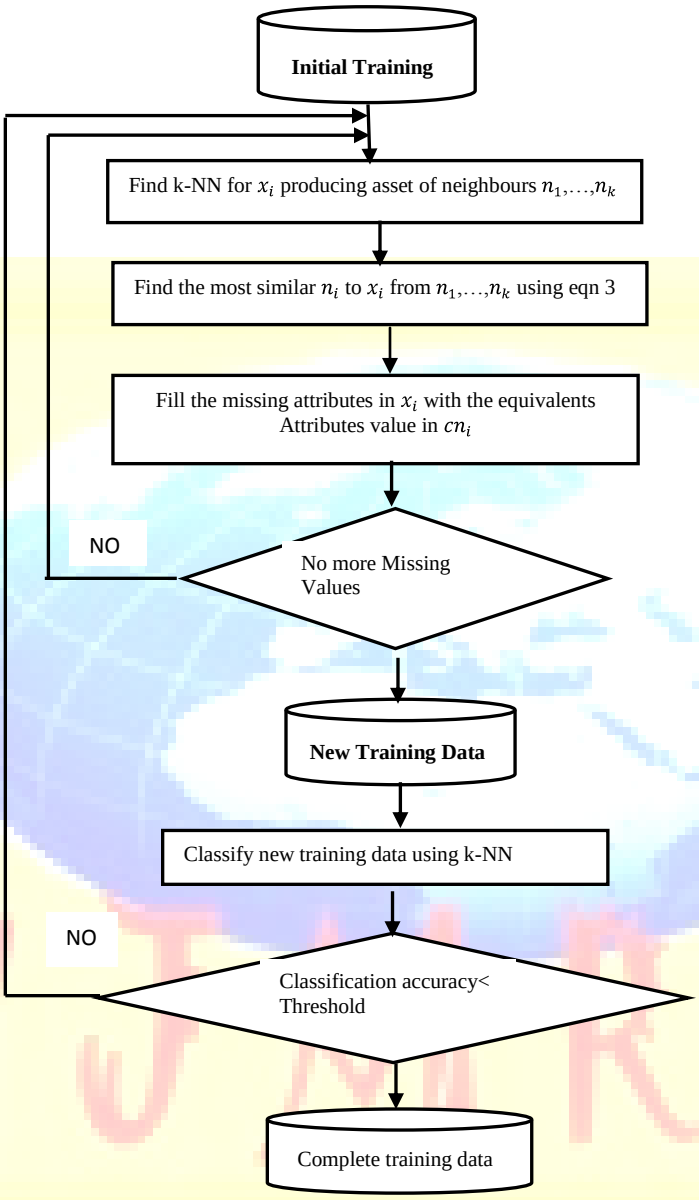

Figure 1: The Flowchart for the proposed method
(Constructing Missing Features Values

The experiment of constructing the missing feature values using the proposed method through *k*-NN classifier with the Minkowski distance function showed that classification accuracy is enhanced by 0.005 when *k*=3 and r=1.5 from the first iteration and a maximum classification accuracy of 0.9698 . Figure 23 shows a comparison of classification accuracy when the missing features values were not treated and when treated using Minkowski/*k*-NN distance metrics.

Further we enhanced our experiment by using some various other methods for distance metrics. However the experiment 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.

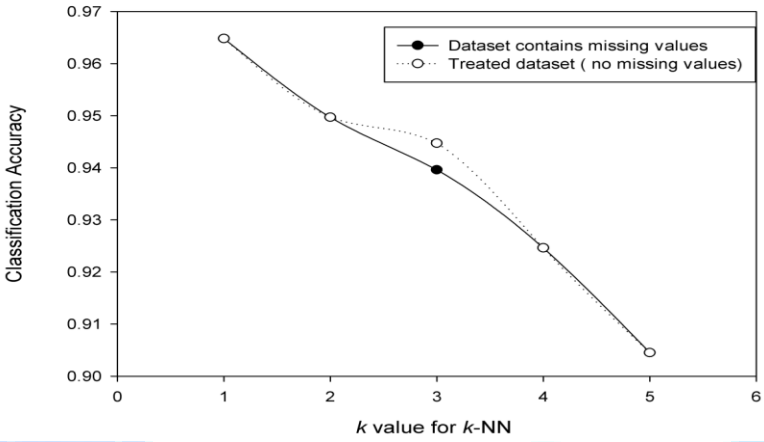

Figure 2: A comparison of classification accuracy for our method through Euclidean/*k*-N

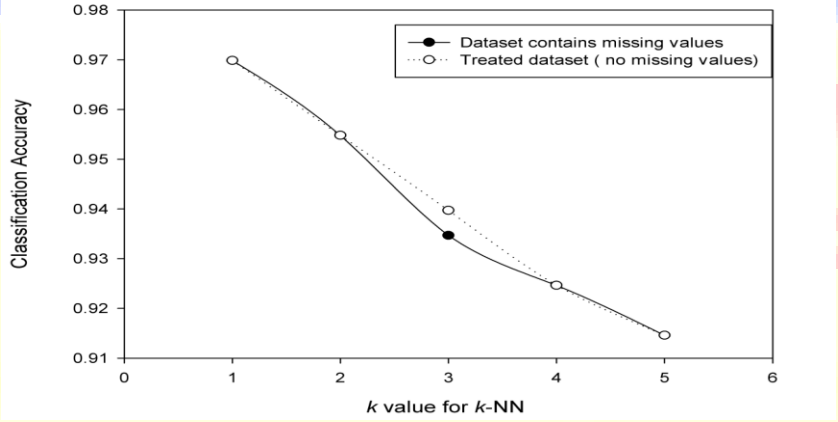

Figure 3: A comparison of classification accuracy for our method through Minkowski/*k*-NN

6. CONCLUSIONS AND FUTURE WORK

Large databases that are used in the medical sector still have a concern of Missing features values that are brought about by many factors as discussed early. This therefore prompted us to propose an approach for constructing missing features values based on iterative *k*-nearest neighbours and the distance functions.

A new approach for constructing missing features values based on iterative k nearest neighbours and the distance functions was proposed. The approach loops until it finds the most suitable feature values that gratify classification accuracy. In consideration of the original dataset, proposed method showed progress of 0.005% in classification accuracy on both Euclidean and Minkowski distance functions. However Manhattan, Chebychev, and Canberra distance metrics produced lower classification accuracy on the new dataset than the original dataset. Accordingly, it was noticed that classification accuracy depended to a great extent on the number of neighbours (k). Investigational evaluation indicated that the less the number of neighbours the more the accuracy. The maximum accuracy of 0.9698 was found when k=1. Further study showed lower classification accuracy on the new dataset than the original dataset when using Manhattan, Chebychev, and Canberra distance functions. Classification accuracy according to this study depended greatly upon the number of neighbours (*k*). To be specific; the maximum classification accuracy was on *k*=1 which was 0.9698 which also can be rephrased as; the less the number of neighbours, the more the classification accuracy and vice versa. In my own opinion, this was brought about by the amount of noise that is produced from conflicting neighbours.

ACKNOWLEDGMENT

This proposal is as a result of inspirational assistance and guidance of my Dad, mentors, lecturers, professionals, and the administrative staff at the Jomo Kenyatta University of Agriculture and Technology as well as at the Machakos university college, my work place. First and foremost, I am grateful to my supervisors; Dr. George Okeyo and Dr. Wilson Cheruiyot, for their invaluable and continuous guidance during the conception of this paper. My other regards goes Dr. Kimwele, who has always kept me on toes and encourage on this research study. They have all given me an undisputed considerable help in every way possible.

I owe you all!

REFERENCES

1. Young, M., et al., *Distance Metrics Overview. 2013. Available from: http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Distance_Metrics_Overview.htm.*
2. Odeh S. et al., *A computer aided diagnosis systems: using genetic algorithm with classifier of the k-nearest neighbors, The International Arab Conference on Information Technology, Tunis.* 2008.
3. World Health Organization Assesses the World's Health Systems. World Health Organization, 2010 website: http://www.who.int/whr/2000/media_centre/press_release/en/index.html.
4. Most Frequent Cancers in Men and Women. 2012; website: http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900.
5. Priddy and Keller (2005): *Artificial neural networks, an introduction*. Washington: SPIE.
6. Kotsiantis, S., Supervised Machine Learning: *a Review of Classification Techniques. Informatica, 2007. 31: p. 249-268.*
7. Daniel T. Larose (2013). "*Discovering Knowledge in Data"* Uniqueness of medical data mining. *Artif. Intell. Med.,* 2013. 26(1-2): p. 1-24.
8. Information on UCI Machine Learning Repository - the *website: http://archive.ics.uci.edu/ml/about.html.*
9. Marlin, B., Missing Data Problems in Machine Learning, in Department of Computer Science. 2008, University of Toronto: Canada
10. Howell D. (2009) "*Treatment of Missing Data".*
11. Rubin, D.B., Inference and missing data. *Biometrika,* 1976. 63(3): p. 581-592.
12. Grzymala-Busse, J.W. and W.J. Grzymala-Busse, *Handling Missing Attribute Values Handbook, O. Maimon and L. Rokach, (Editors). 2010, Springer US. p. 33-51.*
13. Moss, S. Expectation maximization--to manage missing data. 2009.