INTEGRATING NAIVE BAYES AND K-MEANS CLUSTERING WITH DIFFERENT INITIAL CENTROID SELECTION METHODS IN THE DIAGNOSIS OF HEART DISEASE PATIENTS

Mai Shouman[1], Tim Turner[2], Rob Stocker[3]

School of Engineering and Information Technology
University of New South Wales at the Australian Defence Force Academy
Northcott Drive, Canberra ACT 2600
[1] mai_shouman@yahoo.com
[2] t.turner@adfa.edu.au
[3] r.stocker@adfa.edu.au

ABSTRACT

Heart disease is the leading cause of death in the world over the past 10 years. Researchers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease. Naïve Bayes is one of the data mining techniques used in the diagnosis of heart disease showing considerable success. K-means clustering is one of the most popular clustering techniques; however initial centroid selection strongly affects its results. This paper demonstrates the effectiveness of an unsupervised learning technique which is k-means clustering in improving supervised learning technique which is naïve bayes. It investigates integrating K-means clustering with Naïve Bayes in the diagnosis of heart disease patients. It also investigates different methods of initial centroid selection of the K-means clustering such as range, inlier, outlier, random attribute values, and random row methods in the diagnosis of heart disease patients. The results show that integrating k-means clustering with naïve bayes with different initial centroid selection could enhance the naïve bayes accuracy in diagnosing heart disease patients. It also showed that the two clusters random row initial centroid selection method could achieve higher accuracy than other initial centroid selection methods in the diagnosis of heart disease patients showing accuracy of 84.5%.

KEYWORDS

Data Mining, Naïve Bayes, K-Means Clustering, Initial Centroid Selection Methods, Heart Disease Diagnosis.

1. INTRODUCTION

The World Health Organization reported that heart disease is the first leading cause of death in high and low income countries [1]. The European Public Health Alliance reported that heart attacks and other circulatory diseases account for 41% of all deaths [2]. The Economical and Social Commission of Asia and the Pacific reported that in one fifth of Asian countries, most lives are lost to non-communicable diseases such as cardiovascular, cancers, and diabetes diseases [3]. Statistics of South Africa reported that heart and circulatory system diseases are the third leading cause of death in Africa [4]. The Australian Bureau of Statistics reported that heart and circulatory system diseases are the first leading cause of death in Australia, causing 33.7% all deaths [5].

Sundarapandian et al. (Eds): ICAITA, SAI, SEAS, CDKP, CMCA, CS & IT 08,
pp. 125–137, 2012. © CS & IT-CSCP 2012 DOI : 10.5121/csit.2012.2511

Motivated by the world-wide increasing mortality of heart disease patients each year, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease [6-7]. Data mining is an essential step in knowledge discovery. It is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to be detected with traditional statistical methods [8-12]. The application of data mining is rapidly spreading in a wide range of sectors such as analysis of organic compounds, financial forecasting, weather forecasting and healthcare [13].

Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases [14]. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes [15], stroke [16], cancer [17], and heart disease [18]. Several data mining techniques are used in the diagnosis of heart disease such as Naïve Bayes, Decision Tree, neural network, kernel density, bagging algorithm, and support vector machine showing different levels of accuracies [18-24].

Naïve Bayes is one of the successful data mining techniques used in the diagnosis of heart disease patients [22-23]. Although researchers are investigating enhancing naïve bayes performance in classification problems [25], less research is done on enhancing naïve bayes performance in disease diagnosis. This research investigates enhancing naïve bayes performance in the diagnosis of heart disease patients through integrating clustering as a pre-processing step to naïve bayes classification.

K-means clustering is one of the most popular and well know clustering techniques. Its simplicity and good behaviour made it popular in many applications [26]. Initial centroid selection is a critical issue in k-means clustering and strongly affects its results [27]. This paper investigates integrating k-means clustering using different initial centroid selection methods with naïve bayes in the diagnosis of heart disease patients. The rest of the paper is divided as follows: the background section investigates applying data mining techniques in the diagnosis of heart disease; the methodology section explains k-means clustering, different initial centroid selection methods; and naïve bayes used in the diagnosis of heart disease patients; the heart disease data section explains the data used, the results section presents the results of integrating k-means clustering and naïve bayes; followed by the summary section.

2. BACKGROUND

Researchers have been investigating the use of statistical analysis and data mining techniques to help healthcare professionals in the diagnosis of heart disease. Statistical analysis has identified the risk factors associated with heart disease to be age, blood pressure, smoking [28], cholesterol [29], diabetes [30], hypertension, family history of heart disease [31], obesity, and lack of physical activity [32]. Knowledge of the risk factors associated with heart disease helps health care professionals to identify patients at high risk of having heart disease.

Researchers have been applying different data mining techniques over different heart disease datasets to help health care professionals in the diagnosis of heart disease [18-19, 22-24, 33]. The results of the different data mining research cannot be compared because they have used different datasets. However, over time a benchmark data set has arisen in the literature: the Cleveland Heart Disease Dataset (CHDD).1

Naïve bayes is one of the data mining techniques showing considerable success compared to other data mining techniques over different heart disease datasets [19, 21-23, 34]. Palaniappan

[1] http://archive.ics.uci.edu/ml/datasets/Heart+Disease.

and Awang investigated comparing different data mining techniques in the diagnosis of heart disease patients. These techniques involved naïve bayes, decision tree, and neural network. The results showed that the naïve bayes could achieve the best accuracy in the diagnosis of heart disease patients [34]. Rajkumar and Reena investigated comparing naïve bayes, k-nearest neighbour, and decision list in the diagnosis of heart disease patients. The results showed that the naïve bayes could achieve the best accuracy in the diagnosis of heart disease patients [21]. Applying naïve bayes in diagnosing heart disease patients showed different accuracies on different datasets that ranged between 62% and 95% [22, 34]. Cheung applied naive bayes classifier on the Cleveland heart disease dataset showing accuracy of 81.48% [35].

Researchers are investigating enhancing naïve bayes performance in classification problems. Ratanamahatana and Gunopulos described a selective bayesian classifier that uses only the features that C4.5 would use in its decision tree showing that selective bayesian classifier performs reliably better than naïve bayes on the ten tested datasets [36]. Ramana, Babu et al. investigated integrating the naïve bayes classification technique with bagging and boosting in the diagnosis of Liver diseases to enhance the performance of the naïve bayes classifier [25]. Although researchers are investigating enhancing naïve bayes performance, less research is done on enhancing naïve bayes performance in the diagnosis of heart disease patients. This paper demonstrates the effectiveness of unsupervised learning such as k-means clustering in improving supervised learning which is naïve bayes in the diagnosis of heart disease patients.

K-means clustering is one of the most popular clustering techniques; however initial centroid selection is a critical issue that strongly affects its results. This paper investigates applying different methods of initial centroid selection such as range, inlier, outlier, random attribute values, and random row methods for k-means clustering technique in the diagnosis of heart disease patients. This paper investigates if integrating k-means clustering with naïve bayes can enhance the classifier's performance in diagnosing heart disease patients. Importantly, the research involves a systematic investigation of which initial centroid selection method can provide better performance in diagnosing heart disease patients. It also investigates if applying different numbers of clusters can provide different performance in diagnosing heart disease patients and which number of clusters will provide the better performance.

3. METHODOLOGY

The methodology section discusses k-means clustering with five initial centroid selection methods. It also discusses the naïve bayes classifier used in the diagnosis of heart disease patients (Figure 1).

Naïve bayes cannot deal with continuous attributes so they need to be converted into discrete ones, a process called discretization. Dougherty et al. carried out a comparative study between two unsupervised and two supervised discretization methods using 16 data sets showing that differences between the classification accuracies achieved by different discretization methods are not statistically significant [37]. Equal frequency discretization is a popular and successful unsupervised discretization method [38]. Previous related research has shown that this discretization method provides marginally better accuracy when applied on the CHHD. So it is used as a pre-processing step to convert the continuous heart disease attributes to discrete ones.

3.1 K-Means Clustering

K-means clustering is one of the most popular and well know clustering techniques because of its simplicity and good behaviour in many applications [26, 38]. The steps used in k-means clustering are shown in Figure 1.

Several researchers have identified that age, blood pressure and cholesterol are critical risk factors associated with heart disease [28, 31-32] . In identifying the attributes that will be used in the clustering, these attributes are obvious clustering attributes for heart disease patients. The number of clusters used in the k- means in this investigation ranged between two and five clusters. The difference between the initial centroid methods is discussed in the following section.

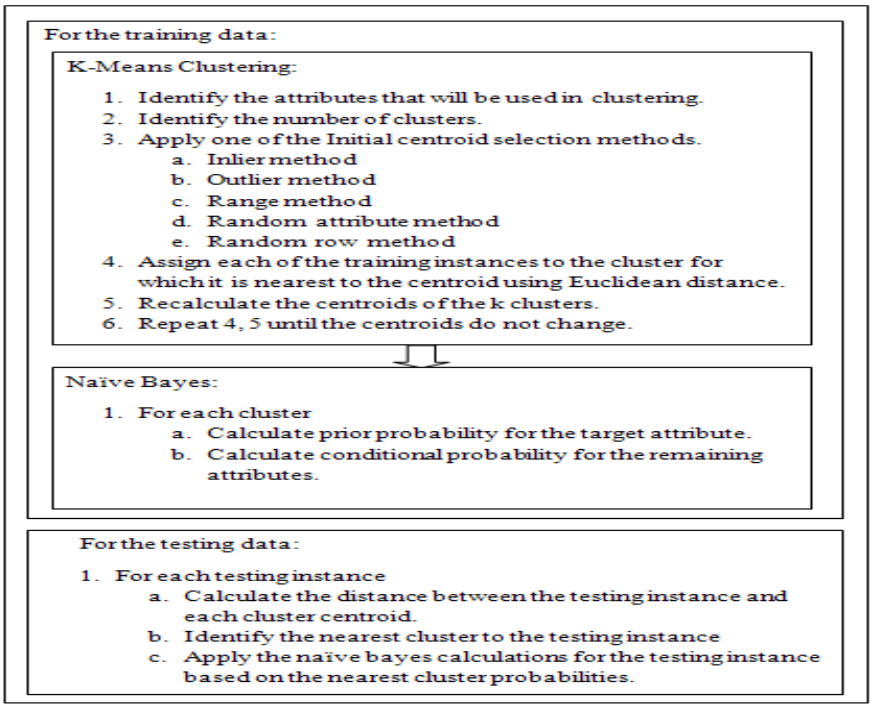

Figure 1: Integrating K-means Clustering and Naïve Bayes

3.2 Initial Centroid Selection

Initial centroid selection is an important matter in k-means clustering and strongly affects its results [27]. This section discusses the generation of initial centroids based on actual sample data points using inlier method, outlier method, range method, random attribute method, and random row method [39]

3.2.1 Inlier Method

In generating the initial K centroids using inlier method the following equations are used:

$C_i = \text{Min}(X) - i$ where $0 \leq i \leq k$ (1)

$C_j = \text{Min}(Y) - j$ where $0 \leq j \leq k$ (2)

Where the initial centroid is C (c_i, c_j) and min (X) and min (Y) are the minimum value of attribute X, and attribute Y respectively. K represents the number of clusters.

3.2.2 Outlier Method

In generating the initial K centroids using outlier method the following equations are used:

$C_i = \text{Max}(X) - i$ where $0 \leq i \leq k$ (3)

$C_j = \text{Max}(Y) - j$ where $0 \leq j \leq k$ (4)

Where the initial centroid is C (c_i, c_j) and max (X) and max (Y) are the maximum value of attribute X, and attribute Y respectively.

3.2.3 Range Method

In generating the initial K centroids using range method the following equations are used:

$C_i = ((Max\ (X) - Min\ (X)) / K) * n$ where $0 \leq i \leq k$ (5)

$C_j = ((Max\ (Y) - Min\ (Y)) / K) * n$ where $0 \leq j \leq k$ (6)

The initial centroid is C (c_i, c_j). Where max (X) and min (X) are maximum and minimum values of attribute X, max (Y) and min (Y) are maximum and minimum values of attribute Y respectively.

3.2.4 Random Attribute Method

In generating the initial K centroids using random attribute method the following equations are used:

$C_i = random(X)$ where $1 \leq i \leq k$ (7)

$C_j = random(Y)$ where $1 \leq j \leq k$ (8)

The initial centroid is C (c_i, c_j). The values of 'i', and 'j' vary from 1 to 'k'.

3.2.5 Random Row Method

In generating the initial K centroids using random row method the following equations are used:

$I = random\ (V)$ where $1 \leq V \leq N$ (9)

$C_i = X\ (I)$ (10)

$C_j = Y\ (I)$ (11)

The initial centroid is C (c_i, c_j). N is the no of instances in the training dataset. X (I) and Y(I) are the values of the attributes X and Y respectively for the instance I.

3.3 Naïve Bayes

Naïve bayes is one of the data mining techniques that show considerable success in classification problems and specially in diagnosing heart disease patients [19, 22]. Naïve bayes is based on probability theory to find the most likely possible classifications [26]. It is based on prior probability of the target attribute and the conditional probability of the remaining attributes. For the training data the prior and conditional probability are calculated for each cluster. For each testing instance in the testing dataset, the probability is calculated with each of the target attribute values and the target attribute value with the largest probability is then selected. The probability of the testing instance for the target attribute value is calculated using the following formula:

$$P\ (v=ci) = P(ci) \times \sum_{j=1}^{n} P(aj = vj \mid class = ci) \quad (12)$$

Where v is the testing instance, ci is the target attribute value, aj is a data attribute and vj is its value [38].

3.4 10 Fold Cross Validation

To measure the stability of the performance of the proposed model the data is divided into training and testing data with 10-fold cross validation. The sensitivity, specificity, and accuracy are calculated. The sensitivity is the proportion of positive instances that are correctly classified as positive (e.g. the proportion of sick people that are classified as sick). The specificity is the proportion of negative instances that are correctly classified as negative (e.g. the proportion of

healthy people that are classified as healthy). The accuracy is the proportion of instances that are correctly classified [38].

Sensitivity = True Positive / Positive (13)

Specificity = True Negative / Negative (14)

Accuracy = (True Positive + True Negative) / (Positive + Negative) (15)

4. HEART DISEASE DATA

The data used in this study is the Cleveland Clinic Foundation Heart disease data set available at *http://archive.ics.uci.edu/ml/datasets/Heart+Disease*. The data set has 76 raw attributes. However, all of the published experiments only refer to 13 of them. Consequently, to allow comparison with the literature, we restricted testing to these same attributes (see Table 1). The data set contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment.

Table 1: Selected Cleveland Heart Disease Data Set Attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	1 = male 0 = female
Cp	Discrete	Chest pain type: 1 = typical angina 2 = atypical angina 3 = non-anginal pain 4 =asymptomatic
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar > 120 mg/dl: 1 = true 0 = false
Restecg	Discrete	Resting electrocardiographic results: 0 = normal 1 = having ST-T wave abnormality 2 =showing probable or define left ventricular hypertrophy by Estes'criteria
Thalach	Continuous	Maximum heart rate achieved
Exang	Discrete	Exercise induced angina: 1 = yes 0 = no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment : 1 = up sloping 2 = flat 3= down sloping
Ca	Discrete	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6= fixed defect 7= reversible defect
Diagnosis	Discrete	Diagnosis classes: 0 = healthy 1= patient who is subject to possible heart disease

5. RESULTS

The results of sensitivity, specificity, and accuracy in the diagnosis of heart disease using k-means clustering and naïve bayes with different initial centroids selection methods and different numbers of clusters are shown in Table 2. For the random attribute and random row methods, ten runs are executed and the average and best for each method are calculated and shown in Table 2.

Table 2: Integrating different initial centroid selection for k-means clustering with naïve bayes in diagnosing heart disease patients

No of Clusters	Initial Centroid Selection Method		Sensitivity	Specificity	Accuracy
No of clusters = 2	Inlier Method		74.1	79.9	83.2
	Outlier Method		74.3	80.9	83
	Range Method		74.3	80.9	83
	Random Attribute	Avg	74.01	79.82	82.56
		Best	73.3	82.5	84.2
	Random Row	Avg	74.31	80.43	83.12
		Best	**75.7**	**81.9**	**84.5**
No of clusters = 3	Inlier Method		74.9	81.1	83.8
	Outlier Method		74.9	81.1	83.8
	Range Method		78.2	78.3	84.2
	Random Attribute	Avg	73.6	79.4	82.37
		Best	74.7	81.2	83.4
	Random Row	Avg	74.36	79.41	82.48
		Best	77.6	79.1	84
No of clusters = 4	Inlier Method		77.4	77.1	83.2
	Outlier Method		70.3	79.3	80.2
	Range Method		77.4	77.1	83.2
	Random Attribute	Avg	73.05	77.21	80.9
		Best	77.8	76.5	82.6
	Random Row	Avg	73.84	77.07	81.11
		Best	75.3	78.5	82.6
No of clusters = 5	Inlier Method		72	71.8	77.9
	Outlier Method		70.9	79.8	81.6
	Range Method		72	71.8	77.9
	Random Attribute	Avg	71.66	75.69	79.23
		Best	70.4	78.7	80.3
	Random Row	Avg	69.93	76.59	78.62
		Best	72.5	77.8	79.9

Table 2 shows that there is no significant difference between the accuracy of two clusters of the inlier, outlier, and range initial centroid selection methods in the diagnosis of heart disease patients. It also shows that the random attribute and random row methods achieved higher accuracy than inlier, outlier, and range methods with two clusters. The best accuracy achieved is by two clusters random row initial centroid selection method showing accuracy of 84.5% as shown in Table 2. The sensitivity, specificity, and accuracy achieved by increasing the number of clusters of inlier, outlier, range, random attribute and random row initial centroid selection methods are shown in Figure 2 to 6 respectively.

The number of clusters of the inlier, outlier, and range initial centroid selection methods could enhance their accuracy in the diagnosis of heart disease patients showing the best accuracy with the three clusters. However increasing the number of clusters more than three clusters decreases their accuracy in the diagnosis of heart disease patients. Also the increase shown in the accuracy with the three clusters of the inlier, outlier, and range initial centroid selection methods is still less than that achieved by two clusters random row initial centroid selection as shown in Figure 7. Increasing the number of clusters of the random attribute and random row initial centroid selection methods did not show any enhancement in their accuracy in the diagnosis of heart disease patients as shown in Figure 7.

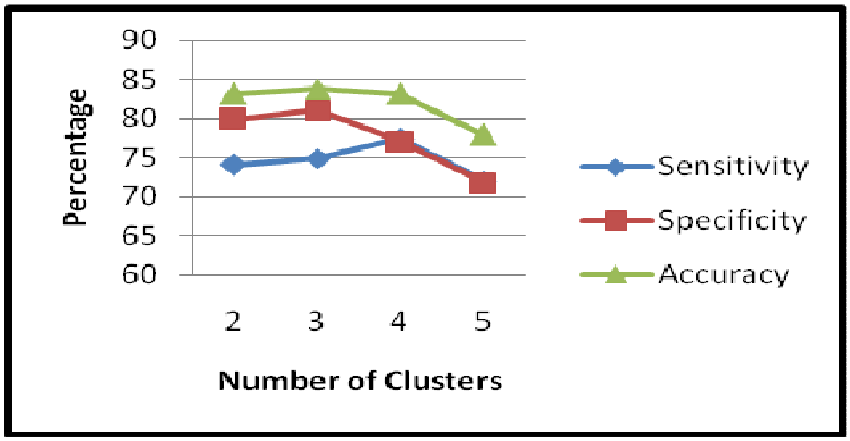

Figure 2: Different Number of Clusters Performance for Inlier Method

Figure 3: Different Number of Clusters Performance for Outlier Method

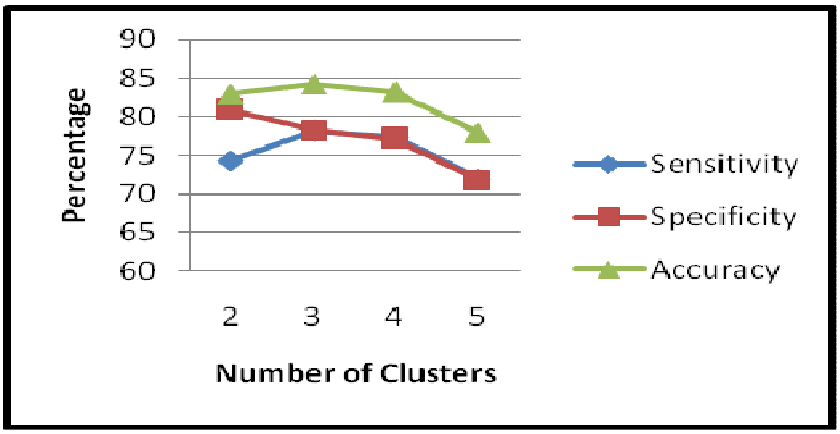

Figure 4: Different Number of Clusters Performance for Range Method

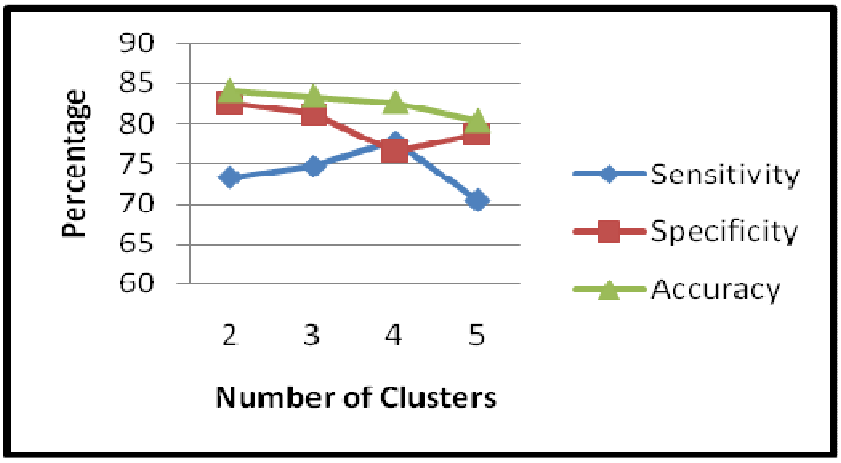

Figure 5: Different Number of Clusters Performance for Random Attribute Method

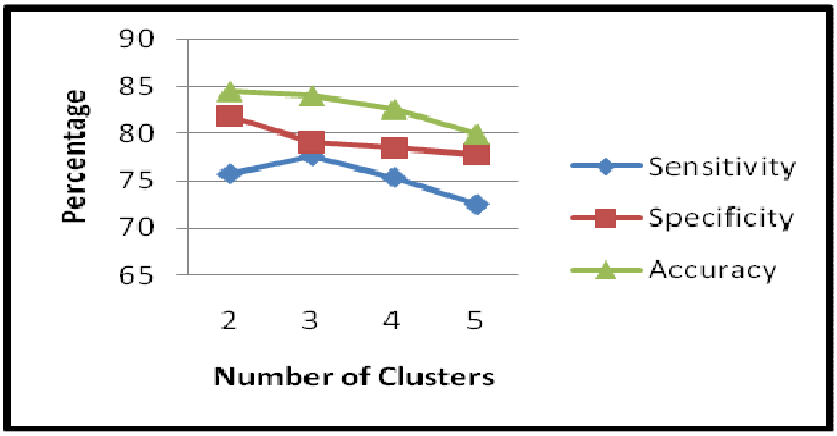

Figure 6: Different Number of Clusters Performance for Random Row Method

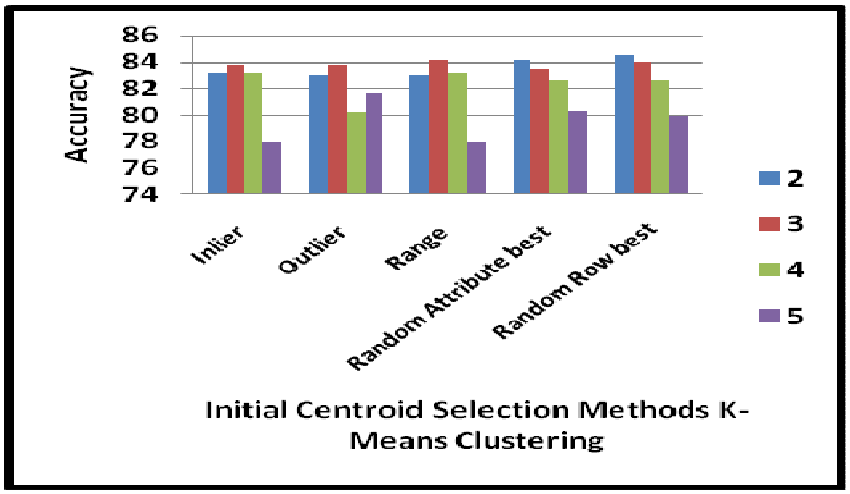

Figure 7: Different Number of Clusters Accuracy with Different Initial Centroid Selection Methods

When comparing integrating k-means clustering and naïve bayes with traditional naïve bayes applied previously on the same dataset, integrating k-means clustering with naïve bayes could enhance the accuracy of naïve bayes in diagnosing heart disease patients as shown in Table 3. Also integrating k-means clustering and naïve bayes could achieve higher accuracy than the decision tree and bagging algorithm in the diagnosis of heart disease patients as shown in Table 3.

Table 3: Comparing integrating k-means clustering and naïve bayes with traditional naïve bayes and other data mining techniques

Author/ Year	Technique	Accuracy
Cheung, N., 2001	Naïve Bayes	81.48%
Tu, et al., 2009	Decision tree	78.91%
	Bagging Algorithm	81.41%
Our work	Two Clusters Random Row Initial Centroid Selection K-Means Clustering with Naïve Bayes	84.5%

Integrating k-means clustering with naïve bayes showed the best accuracy with the Random Row two clusters k-means clustering naïve bayes. Why do two clusters show better performance than other numbers of clusters in the diagnosis of heart disease patients? There is a need for an explanation why the two clusters showed better performance than other numbers of clusters in the diagnosis of heart disease patients. The number of instances is relatively small in the CHHD. A larger dataset is needed to identify if two clusters will still provide the best accuracy results. Also, the target attribute of the Cleveland heart disease dataset has two values which are health and sick. Further investigation is also needed to identify if there is a relationship between the number of clusters showing best accuracy results and the number of values of the target attribute. Although a larger data set is needed to verify the results of integrating naïve bayes and k-means clusters; however the Cleveland heart disease data is used as an initial step to be able to compare its results with other data mining techniques results previously applied on the same data set. As a

future work investigating integrating naïve bayes and k-means clusters will be applied on a larger data set obtained from heart disease hospital.

6. SUMMARY

Heart disease is the leading cause of death all over the world in the past ten years. Researchers have been investigating applying different data mining techniques to help health care professionals in the diagnosis of heart disease. Naïve bayes is one of the successful data mining techniques used in the diagnosis of heart disease patients. This paper investigates integrating k-means clustering with naïve bayes in the diagnosis of heart disease patients. Initial centroid selection is a critical issue that strongly affects k-means clustering results. This paper systematically investigates applying different methods of initial centroid selection such as range, inlier, outlier, random attribute values, and random row methods for k-means clustering technique in the diagnosis of heart disease patients. The results show that integrating k-means clustering and naïve bayes can enhance naïve bayes accuracy in the diagnosis of heart disease patients. The results also show that the random attribute and random row methods could achieve higher accuracy than inlier, outlier, and range methods in the diagnosis of heart disease patients. The best accuracy achieved is by two clusters random row initial centroid selection method showing accuracy of 84.5%. Finally, some limitations on this work are noted as pointers for future research.

REFERENCES

[1] World Health Organization. 2007 7-Febuary 2011]; Available from: http://www.who.int/mediacentre/factsheets/fs310.pdf.

[2] European Public Health Alliance. 2010 7-February-2011]; Available from: http://www.epha.org/a/2352

[3] ESCAP. 2010 7-February-2011]; Available from: http://www.unescap.org/stat/data/syb2009/9.Health-risks-causes-of-death.asp.

[4] Statistics South Africa. 2008 7-February-2011]; Available from: http://www.statssa.gov.za/publications/P03093/P030932006.pdf

[5] Australian Bureau of Statistics. 2010 7-February-2011]; Available from: http://www.ausstats.abs.gov.au/Ausstats/subscriber.nsf/0/E8510D1C8DC1AE1CCA2576F600139288/$File/33030_2008.pdf

[6] Helma, C., E. Gottmann, and S. Kramer, Knowledge discovery and data mining in toxicology. Statistical Methods in Medical Research, 2000.

[7] Podgorelec, V., et al., Decision Trees: An Overview and Their Use in Medicine. Journal of Medical Systems, 2002. Vol. 26.

[8] Han, j. and M. Kamber, Data Mining Concepts and Techniques. 2006: Morgan Kaufmann Publishers.

[9] Lee, I.-N., S.-C. Liao, and M. Embrechts, Data mining techniques applied to medical information. Med. inform, 2000.

[10] Obenshain, M.K., Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 2004.

[11] Sandhya, J., et al., Classification of Neurodegenerative Disorders Based on Major Risk Factors Employing Machine Learning Techniques. International Journal of Engineering and Technology, 2010. Vol.2,[No.4.

[12] Thuraisingham, B., A Primer for Understanding and Applying Data Mining. IT Professional IEEE, 2000.

[13] Ashby, D. and A. Smith, The Best Medicine? Plus Magazine - Living Mathematics., 2005.

[14] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classification techniques. MED. INFORM., 2002. Vol. 27, no. 1, 59–67, .

[15] Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Americas Conference on Information Systems, 2009.

[16] Panzarasa, S., et al., Data mining techniques for analyzing stroke care processes. Proceedings of the 13th World Congress on Medical Informatics, 2010.

[17] Li L, T.H., Wu Z, Gong J, Gruidl M, Zou J, Tockman M, Clark RA, Data mining techniques for cancer detection using serum proteomic profiling. Artificial Intelligence in Medicine, Elsevier, 2004.

[18] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009. 36 (2009): p. 7675–7680.
[19] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.
[20] Hara, A. and T. Ichimura, Data Mining by Soft Computing Methods for The Coronary Heart Disease Database. Fourth International Workshop on Computational Intelligence & Applications, IEEE, 2008.
[21] Rajkumar, A. and G.S. Reena, Diagnosis Of Heart Disease Using Datamining Algorithm. Global Journal of Computer Science and Technology, 2010. Vol. 10 (Issue 10).
[22] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.
[23] Srinivas, K., B.K. Rani, and A. Govrdhan, Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks. International Journal on Computer Science and Engineering (IJCSE), 2010. Vol. 02, No. 02: p. 250-255.
[24] Yan, H., et al., Development of a decision support system for heart disease diagnosis using multilayer perceptron. Proceedings of the 2003 International Symposium on, 2003. vol.5: p. pp. V-709- V-712.
[25] Ramana, B.V., M.S.P. Babu, and N.B. Venkateswarlu, A critical evaluation of bayesian classifier for liver diagnosis using bagging and boosting methods. International Journal of Engineering Science and Technology, 2011. Vol. 3 No. 4.
[26] Wu, X., et al., Top 10 algorithms in data mining analysis. Knowl. Inf. Syst., 2007.
[27] Tajunisha, N. and V. Saravanan, A new approach to improve the clustering accuracy using informative genes for unsupervised microarray data sets. International Journal of Advanced Science and Technology, 2011.
[28] Heller, R.F., et al., How well can we predict coronary heart disease? Findings in the United Kingdom Heart Disease Prevention Project. BRITISH MEDICAL JOURNAL, 1984.
[29] Wilson, P.W.F., et al., Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association Journal, 1998.
[30] Simons, L.A., et al., Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia, 2003. 178.
[31] Salahuddin and F. Rabbi, Statistical Analysis of Risk Factors for Cardiovascular disease in Malakand Division. Pak. j. stat. oper. res., 2006. Vol.II: p. pp49-56.
[32] Shahwan-Akl, L., Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 2010. 6 (1).
[33] Tu, M.C., D. Shin, and D. Shin, Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics, IEEE, 2009.
[34] Palaniappan, S. and R. Awang, Web-Based Heart Disease Decision Support System using Data Mining Classification Modeling Techniques. Proceedings of iiWAS, 2007.
[35] Cheung, N., Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queenland., 2001.
[36] Ratanamahatana , C.A. and D. Gunopulos, Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection. Proc. Workshop Data Cleaning and Preprocessing (DCAP '02), at IEEE Int'l Conf. Data Mining (ICDM '02), 2002.
[37] Dougherty, J., R. Kohavi, and M. Sahami, Supervised and unsupervised discretization of continuous features. In: Proceedings of the 12th international conference on machine learning. San Francisco: Morgan Kaufmann, 1995: p. p. 194–202.
[38] Bramer, M., Principles of data mining. 2007: Springer.
[39] Khan, D.M. and N. Mohamudally, A Multiagent System (MAS) for the Generation of Initial Centroids for kmeans Clustering Data Mining Algorithm based on Actual Sample Datapoints. Journal of Next Generation Information Technology, 2010. Volume 1, Number 2.

Authors

1. Mrs Mai Mohammed Shouman: PhD Candidate Lecturer in School of Engineering and Information Technology UNSW@CANBERRA

Email: mai.shouman@student.adfa.edu.au

2. Dr Tim Turner: Senior Lecturer in School of Engineering and Information Technology UNSW@CANBERRA

Email: t.turner@adfa.edu.au

3. Dr Robert Stocker: Visiting Fellow in School of Engineering and Information Technology UNSW@CANBERRA

Email : r.stocker@adfa.edu.au