2012 International Conference on Knowledge Discovery (ICKD 2012)

IPCSIT vol. XX (2012) © (2012) IACSIT Press, Singapore

Applying k-Nearest Neighbour in Diagnosing Heart Disease Patients

Mai Shouman[1*], Tim Turner[1], Rob Stocker[1]

[1]School of Engineering and Information Technology
University of New South Wales at the Australian Defence Force Academy
Northcott Drive, Canberra ACT 2600

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. K-Nearest-Neighbour (KNN) is one of the successful data mining techniques used in classification problems. However, it is less used in the diagnosis of heart disease patients. Recently, researchers are showing that combining different classifiers through voting is outperforming other single classifiers. This paper investigates applying KNN to help healthcare professionals in the diagnosis of heart disease. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The results show that applying KNN could achieve higher accuracy than neural network ensemble in the diagnosis of heart disease patients. The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease.

Keywords: Data Mining, K-Nearest-Neighbour, Voting, Heart Disease.

1. Introduction

Heart disease is the leading cause of death in the world over the past 10 years. 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 Economic 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]. 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 [4].

Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease [5-6]. 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 [7-11]. The application of data mining is rapidly spreading in a wide range of sectors such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting [12].

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 [13]. Researchers are using data mining techniques in the medical diagnosis of several diseases such as diabetes [14], stroke [15], cancer [16], and heart disease [17]. Several data mining techniques are used in the diagnosis of heart disease showing different levels of accuracy.

K-Nearest-Neighbour (KNN) is one of the most widely used data mining techniques in pattern recognition and classification problems [18]. Recently Paris et al. examined single classifiers and combining different classifiers through voting and showed that voting outperformed other single classifiers [19]. This paper investigates applying KNN in the diagnosis of heart disease on the benchmark dataset to allow comparisons with other data mining techniques used on the same dataset. It also investigates if integrating

[*] Corresponding Author. Tel: +61 2 6268 8034 Fax: +61 2 6268 8581 m.shouman@adfa.edu.au

voting with KNN can enhance its accuracy 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 KNN and integrating voting with it in diagnosing heart disease patients, the heart disease data section explains the data used, the results section presents the KNN and voting results, followed by the summary section.

2. Background

Healthcare professionals store significant amounts of patients' data that could be used to extract useful knowledge. 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 habit [20], total cholesterol [21], diabetes [22], hypertension, family history of heart disease [23], obesity, and lack of physical activity [24]. 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 such as decision tree, naïve bayes, neural network, bagging, kernel density, and support vector machine over different heart disease datasets to help health care professionals in the diagnosis of heart disease [17, 25-30]. 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). Results of trials on this dataset do allow comparison. Table 1 shows a sample of data mining techniques used on CHDD in the diagnosis of heart disease patients showing different levels of accuracy that ranged between 81% and 89%. Polat, Sahan et al., used the KNN as a pre-processing step to weight attributes before applying artificial immune recognition system but did not use KNN as a classification technique [31].

Table 1: A Sample of Data Mining Techniques Used on the Cleveland Heart Disease Dataset

Author/ Year	**Technique**	**Accuracy**
(Cheung 2001)	Decision Tree	81.11%
	Naïve Bayes	81.48%
(Polat , Sahan et al. 2007)	Fuzzy-AIRS–K-Nearest Neighbour	87.00%
(Tu, Shin et al. 2009)	Bagging algorithm	81.41%
(Das, Turkoglu et al. 2009)	Neural network ensembles	89.01%
(Shouman, et al 2011)	Nine voting equal frequency discretization gain ratio decision tree	84.10%

K-Nearest-Neighbour is one of the most widely used data mining techniques in classification problems [18]. Its simplicity and relatively high convergence speed make it a popular choice. However a main disadvantage of KNN classifiers is the large memory requirement needed to store the whole sample. When the sample is large, response time on a sequential computer is also large [33]. Despite the memory requirement issue, it is showing good performance in classification problems of various datasets [34].

Recently, researchers are suggesting that applying voting can outperform other single classifiers [19]. Dividing the training data into smaller subsets and building a model for each subset then applying voting to classify testing data can enhance the classifier's performance [35]. Recently, we conducted research on applying Decision Tree and investigated if integrating voting with it can enhance its accuracy in the diagnosis heart disease patients. The results showed that integrating voting with Decision Tree could enhance its accuracy in the diagnosis of heart disease patients. This paper investigates applying KNN to help healthcare professionals in the diagnosis of heart disease. It also investigates if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients.

3. Methodology

The methodology section discusses the voting technique and KNN used in the diagnosis of heart disease patients

3.1. Voting

Voting is an aggregation technique used to combine decisions of multiple classifiers. The idea of applying multiple classifier voting is dividing the training data into smaller equal subsets of data and building a classifier for each subset of data [36]. The simplest form of voting is based on plurality or majority voting, each single classifier contributes a single vote. The final decision is based on the majority of the votes; i.e., the class with the most votes is the final prediction. The final decision is selected by summing up all votes and by choosing the class with the highest aggregate [37]. The number of voting divisions used in this paper ranged between three and eleven subsets.

3.2. K-Nearest-Neighbour (KNN)

KNN is one of the most simple and straight forward data mining techniques. It is called Memory-Based Classification as the training examples need to be in the memory at run-time [33].

When dealing with continuous attributes the difference between the attributes is calculated using the Euclidean distance. If the first instance is (a_1, a_2, a_3 … a_n) and the second instance is (b_1, b_2, b_3, … b_n), the distance between them is calculated by the following formula:

$$\sqrt{(a_1-b_1)^2 + (a_2-b_2)^2 + \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (a_n-b_n)^2} \qquad (1)$$

A major problem when dealing with the Euclidean distance formula is that the large values frequency swamps the smaller ones. For example, in heart disease records the cholesterol measure ranges between 100 and 190 while the age measure ranges between 40 and 80. So the influence of the cholesterol measure will be higher than the age. To overcome this problem the continuous attributes are normalized so that they have the same influence on the distance measure between instances.

KNN usually deals with continuous attributes however it can also deal with discrete attributes. When dealing with discrete attributes if the attribute values for the two instances a_2, b_2 are different so the difference between them is equal to one otherwise it is equal to zero.

3.3. 10 Fold Cross Validation

To evaluate the performance stability 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 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].

3.4. Heart Disease Data

The data used in this study is the benchmark 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. The data set contains 303 rows of which 297 are complete. Six rows contain missing values and they are removed from the experiment.

4. Results

Table 2 shows the sensitivity, specificity, and accuracy results of KNN without voting in the diagnosis of heart disease patients. The value of K ranged between one and thirteen. The accuracy achieved without voting KNN ranged between 94% and 97.4 % with different values of K. The value of K equal to 7 achieved the highest accuracy and specificity (97.4% and 99% respectively).

Apply voting to K-nearest-neighbour did not show any enhancement in the sensitivity, specificity, and accuracy with different values of K. Table 3 shows the results of applying different numbers of subsets voting to KNN at K=7. When applying the voting, the accuracy decreased from 97.4% to 92.7%. Although applying voting to decision tree increased decision tree accuracy, applying voting to KNN did not show any enhancement to its accuracy in diagnosis of heart disease patients as shown in Figure 1.

Table 2: Without Voting K-Nearest Neighbour

	Sensitivity	**Specificity**	**Accuracy**
K=1	93.2%	93.1%	94%
K=3	93.4%	95.7%	96.7%
K=5	93.8%	97.6%	96.7%
K=7	93.8%	99%	97.4%
K=9	95.9%	98.2%	97.3%
K=11	92.5%	98.2%	96.4%
K=13	93.5%	98.7%	97.1%

Table 3: Voting K=7 Nearest Neighbour

	Sensitivity	**Specificity**	**Accuracy**
3 votes	89.8%	96.2%	95%
5 votes	89.5%	98.7%	95.7%
7 votes	91.9%	97.3%	95%
9 votes	83.5%	99%	93%
11 votes	84%	99%	92.7%

So, why does applying voting enhance decision tree accuracy but not enhance KNN accuracy? When applying voting to decision tree, a different decision rules are extracted from each subset, which helps in extracting new knowledge that could increase the decision tree accuracy. In the case of KNN, the distance between the testing instance and different datasets is measured. However, there is no new knowledge extracted, just the distance is measured. Furthermore, partitioning the dataset to develop the different classifiers reduces the range across which the nearest neighbour calculations can reach, potentially reducing the refinement offered from calculations in each partition.

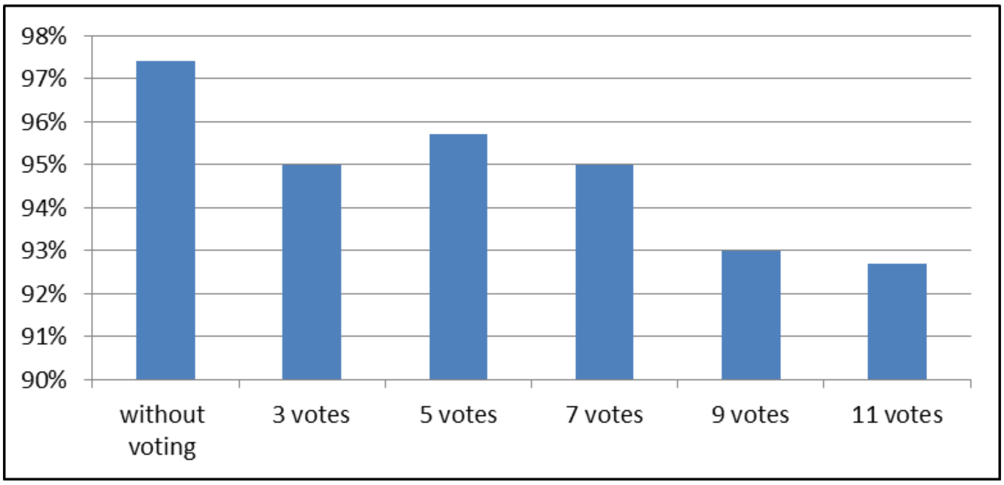

Figure 1: K=7 Nearest Neighbour Accuracy Comparison

When comparing KNN in the diagnosis of heart disease with other data mining techniques used on the same benchmark dataset, the KNN achieved higher accuracy (97.4%) than the highest other published results on the same dataset, which used a neural network ensemble (89.01%) [17].

5. Summary

Heart disease is the leading cause of death all over the world in the past ten years. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patients' data that could be used to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease. In this paper we report research that applied KNN on a benchmark dataset to investigate its efficiency in the diagnosis of heart disease. We also investigated if integrating voting with KNN could enhance its accuracy even further. Our results show that applying KNN achieved an accuracy of 97.4% which is higher than any other published findings on that benchmark dataset. The results also show that applying voting could not enhance the KNN accuracy in the diagnosis of heart disease. Of course, while KNN has produced excellent results, the work needs to be verified against other and larger datasets; work which is ongoing.

6. 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] 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

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

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

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

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

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

[10] 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.

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

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

[13] 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, .

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

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

[16] 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.

[17] 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.

[18] Moreno-Seco, F., L. Mico, and J.A. Oncina, Modification of the LAESA Algorithm for Approximated k-NN Classification. . Pattern Recognition Letters, 2003. 24 p. pp. 47–53.

[19] Paris, I.H.M., L.S. Affendey, and N. Mustapha, Improving Academic Performance Prediction using Voting Technique in Data Mining. World Academy of Science, Engineering and Technology, 2010. 62.

[20] 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.

[21] Wilson, P.W.F., et al., Prediction of Coronary Heart Disease Using Risk Factor Categories. American Heart Association Journal, 1998.

[22] Simons, L.A., et al., Risk functions for prediction of cardiovascular disease in elderly Australians: the Dubbo Study. Medical Journal of Australia, 2003. 178.

[23] 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.

[24] Shahwan-Akl, L., Cardiovascular Disease Risk Factors among Adult Australian-Lebanese in Melbourne. International Journal of Research in Nursing, 2010. 6 (1).

[25] Andreeva, P., Data Modelling and Specific Rule Generation via Data Mining Techniques. International Conference on Computer Systems and Technologies - CompSysTech, 2006.

[26] Sitar-Taut, V.A., et al., Using machine learning algorithms in cardiovascular disease risk evaluation. Journal of Applied Computer Science & Mathematics, 2009.

[27] 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.

[28] Tu, M.C., D. Shin, and D. Shin, Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics, IEEE, 2009.

[29] 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.

[30] Kangwanariyakul, Y., et al., Data mining of magnetocardiograms for prediction of ischemic heart disease. EXCLI Journal, 2010.

[31] Polat , K., S. Sahan, and S. Gunes, Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting preprocessing. Expert Systems with Applications 2007. 32 p. 625–631.

[32] Cheung, N., Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering, B.Sc. Thesis, University of Queenland., 2001.

[33] Alpaydin, E., Voting over Multiple Condensed Nearest Neighbors. Artificial Intelligence Review, 1997. 11: p. 115–132.

[34] Liang-yan, S. and C. Li, A Fast and Scalable Nearest Neighbor Based Classifier for Data Mining. IEEE Global Congress on Intelligent Systems, 2009.

[35] Breiman, L., Pasting bites together for prediction in large data sets. Machine Learning, 1999. vol. 36, no. 1,2, pp. 85-103.

[36] Hall, L.O., et al., Distributed Learning on Very Large Data Sets. In Workshop on Distributed and Parallel Knowledge Discover., 2000.

[37] Paris, I.H.M., L.S. Affendey, and N. Mustapha, Improving Academic Performance Prediction using Voting Technique in Data Mining. World Academy of Science, Engineering and Technology 2010.

[38] Bramer, M., Principles of data mining. 2007: Springer.