Exploring on Various Prediction Model in Data Mining Techniques for Disease Diagnosis

K.Lokanayaki
Assistant Professor
Department of Computer Application
Florence Group of Intuitions
Bangalore

A.Malathi, Ph. D
Assistant Professor
PG and Research Department of Computer Science,
Govt. Arts College
Coimbatore

ABSTRACT

The main objective of this survey paper focused on variety of data mining techniques, approaches and different researches which are ongoing and helpful to medical diagnosis of disease. The survey is conducted in three different dimensions. Study was conducted using classification model, clustering model and bio-inspirational model. The study reveals that depending on the type of dataset used each model differs in their performance. For predicting the disease with labeled dataset the classification model was well suited in that the support vector machine and its variants are highly used. If the dataset consist of unlabelled features then the clustering model better suits for pattern recognition among the several methods k-means algorithm with the improvisation is adapted by researches due to its simplicity. To increase the performance of dataset with more optimization, then the bio-inspirational based techniques is well suited, in this particle swarm optimization is most used because of its bigger optimization ability and it can be completed easily. Thus the paper investigates the importance of each model in the field of medical diagnosis.

Keywords

Medical field, Data Mining Methods, Data mining applications, Prediction, Medical diagnosis.

1. INTRODUCTION

Data Mining plays an important role in the Prediction of Diseases. The data mining methods comparison were targeted as a main objective in many studies that mainly aimed to develop a prediction model in a critical fields, like medicine, by investigating several data mining methods, intending to get the model that have the highest prediction accuracy. The Aim of this survey is to analyze some of the famous earlier works in data mining techniques using classification, clustering and bio-inspirational based models.

In the classification based model the major metrics used for identifying the performance of each classifiers are done using were the Sensitivity, Specificity, Accurateness, Error Rate, TPR and FPR. In clustering based model the performance measures are based on the compactness and separation of clusters and detect the correct number of clusters. The Bio-Inspiration based prediction model mainly concentrates on the low cost, high speed computation with more accuracy rate. The below sections describes the various techniques formulated under these three models to make efficient disease diagnosis.

2. SURVEY ON CLASSIFICATION MODEL BASED PREDICTION

In most of the disease diagnosis problem the classification systems have been used in many cases. The literature survey related to the study of classification relevance are observed, it can be seen that a great range of methods were used which reached high classification accuracies using the disease diagnosis dataset.

The prediction techniques RIPPER, decision tree, neural networks and support vector machine were used to predict cardiovascular disease patients. The performance comparison metrics are done by the True Positive Rate (TPR), False Positive Rate (FPR), Sensitivity, Specificity and Accuracy. The study done by Kumari et al[1] showed that support vector machine model outperforms the other models for predicting cardiovascular disease

To predict the presence of coronary artery disease Imran Kurt et al [2] did comparison performance with logistic regression (LR), classification and regression tree (CART), multi-layer preceptor (MLP), radial basis function (RBF), and self-organizing feature maps (SOFM). The result shows that MLP, CART, LR, and RBF performed better than SOFM in predicting CAD in according to HCA and MDS. Carlos implemented efficient search for diagnosis of heart disease comparing association rules with decision trees [3]. The problem of identifying constrained association rules for heart disease prediction was studied by Carlos. A model intelligent heart diseases prediction system based on decision tree, naïve bayes and neural networks built with the aid of data mining techniques was proposed by sellappan palaniappan et al [4].

3. SURVEY ON CLUSTERING MODEL BASED PREDICTION

Disparate of classification and prediction, which examines class-labeled data objects, clustering investigates data objects exclusive of checking with a recognized class label. In common, the class labels are not present in the training data just because they are not predictable to initiate with. Clustering can be used to engender such labels. The objects are clustered or assembled based on the belief of maximizing the intra-class likeness and minimizing the interclass likeness. That is, clusters of objects are created so that objects within a cluster have elevated likeness in appraisal to one another, but are extremely dissimilar to items in further clusters.

Thangavel et al [5] used the K-means clustering algorithm to investigate cervical cancer patients and initiate that clustering found better prognostic results than existing medical opinion. They initiate a set of motivating attributes that could be used by doctors as supplementary hold up on whether or not to suggest a biopsy for a patient assumed of having the cervical cancer.

Fuzzy cluster means (FCM or fuzzy C-Means) model for the analysis of blood albumin and clinical symptoms to categorize liver disorders. Application of cluster analysis engages a series of procedural and diagnostic decision steps that developed the distinction and consequence of the clusters created. The suspicions often related with exploration of LFT test and clinical data are purged by the proposed system [6]. Wael a. Alzoubi et al [7] proposed scalable and efficient method for mining association rules based On Clustering. In prior work, three dissimilar clustering algorithms are used namely Fuzzy C-Means (FCM) Clustering, Hierarchical Clustering Analysis (HCA), and Simulated Annealing Fuzzy Clustering (SAFC), were examined using data sets that comprised of FTIR spectra collected from oral cancers [8,9,10]. The experimental effect indicated that FCM clustering performed significantly better than HCA when classifying spectra into their explicit diagnosis [14]. It was also shown that when FCM clustering was combined with simulated annealing, the algorithm was capable to involuntarily obtain the best possible number of clusters with respect to the Xie-Beni cluster validity measure [8, 10].

Many clusters validation indices have been developed in the past. In the context of fuzzy methods, some of them only use the membership values of a fuzzy cluster of the data, such as the partition coefficient [11] and partition entropy [12]. The advantage of this type of index is that it is easy to compute but it is only useful for the small number of well-separated clusters. Furthermore, it also lacks direct connection to the geometrical properties of the data. In order to overcome this problem Xie and Beni defined a validity index which measures the compactness and separation of clusters [13]. In this paper, the Xie-Beni index has been chosen as the cluster validity measure because it has been shown to be able to detect the correct number of clusters in several experiments [14].

4. BIO INSPIRATIONAL BASED DISEASE DIAGNOSIS

Chowdhury et al [15] proposed a technique to develop a field programmable gate array (FPGA) based small cost, low power and high speed narrative diagnostic system it is a novel variation of particle swarm optimization called as adaptive perceptive particle swarm optimization has been projected to determine the optimal weights of these pathophysiological parameters for a more accurate diagnosis.

Mona Nagy Elbedwehy et al [16] introduces a new innovative computer-aided analysis system of the heart valve disease using binary particle swarm optimization and support vector machine, along with K-nearest neighbor and with cross-validation leave-one-out method. This approach initiates with binary particle swarm optimization algorithm to make a decision on the most weighted features which is pursue by performing support vector machine to categorize two upshots of the heart signals like whether healthy or containing a heart valve disease, then its classified the presence of a heart valve disease into four types of outcomes.

Artificial immune recognition system has shown an effective performance on several problems such as medical classification problems. Breast cancer and liver disorder are identified by Polat et al [17] and classified using artificial immune recognition system with fuzzy resource allocation method. In [18], to diagnose Pima Indian diabetes using principal component analysis and adaptive neuro-fuzzy inference were adopted. Ganji & al [19] used a fuzzy Ant Colony Optimization; they have reported 79.48% classification accuracy.

In [20], Jayalakshmi et al used the ANN method for diagnosing diabetes, using the Pima Indian diabetes dataset without missing data and obtained 68.56% classification accuracy. Amin Einipour had focused in his work [21] on breast cancer diagnosis by grouping of fuzzy systems and evolutionary algorithms. Ant colony algorithm is employed as evolutionary algorithm to optimize the obtained set of fuzzy rules. The result shows that the proposed approach would be capable of classifying cancer instances with high accuracy rate in addition to adequate interpretability of extracted rules.

5. RESULTS AND DISCUSSIONS

The main goal of this survey paper was to determine how the data mining algorithms are utilized in the existing approaches to overcome the problem of diagnosing diseases in the earlier stages. This paper provides an idea about major life-threatening diseases and their diagnosis using classification, clustering and the bio inspirational based techniques. The future work of our contribution is that instead of just finding the pattern recognition alone the dataset which is used for disease diagnosis has to be enriched with the following process

- Data preprocessing methods to overcome the problem of handling missing values.
- Implementing best feature selection approach to determine the optimistic attributes whose contributes is main in pattern recognition.
- Designing the pattern discovery technique in case of incomplete and vagueness identification in the real time dataset.
- Pruning the rules for generating optimized rules for better detection rate.

The Best Result producing data mining algorithms used for disease diagnosis and prognosis are shown in the figure 1, 2 and 3

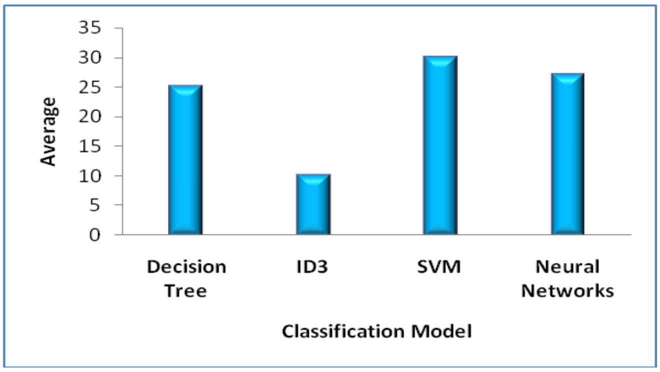

Figure 1 Sample Papers referred for the four algorithms to diagnosis disease based on classification model

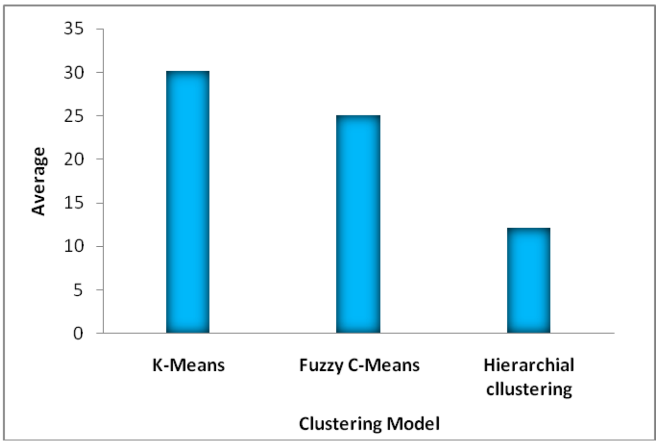

Figure 2 Sample Papers referred for the three algorithms to diagnosis disease based on clustering model.

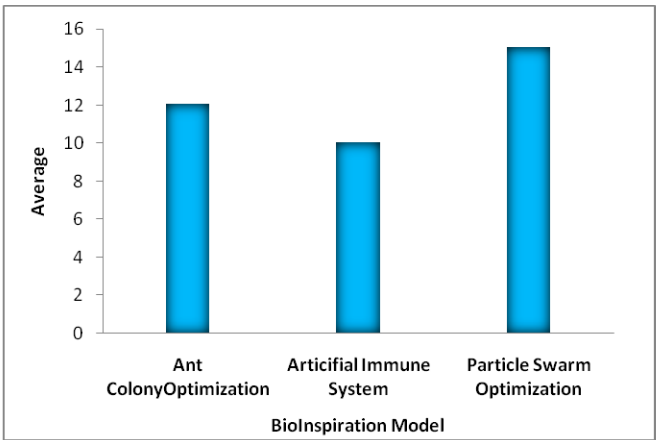

Figure 3 Sample Papers referred for the three algorithms to diagnosis disease based on Bio-Inspiration prediction model.

6. CONCLUSION

Over the past few decades, to automated problem-solving tools, intended to assist the physician in making sense out of the welter of data. In healthcare, data mining is becoming increasingly more essential.

From the above study we observed from more frequent classification models used in disease diagnosis in which support vector machine is used high by many researches because it may be due to it has a regularization parameter, which makes the user think about avoiding over-fitting. It uses the kernel trick, so that it is possible to build in expert knowledge. Thirdly an SVM is defined by a convex optimization problem for which there are efficient methods. Finally, it is estimation to a bounce on the test error rate, and there is a extensive body of theory in the rear it which recommend it should be a good idea.

In the clustering model based forecasting k-means algorithm is used base algorithm and they created variants of k-means to overcome the disadvantages of it. The clusters are non-hierarchical and they do not overlie. Each element of a cluster is nearer to its cluster than any extra cluster because proximity does not always absorb the 'center' of clusters.

In the bio-inspirational based model particle swarm optimization and its variants are most used in the optimization problem of pattern recognition. This is due PSO is based on the intelligence. It can be useful into mutually for technical research and engineering use. PSO have no overlies and mutation computation. The investigation can be carried out by the velocity of the particle. During the progression of several generations, only the most idealist particle can pass on information onto the other particles, and the speed of the re-searching is very rapid. The computation in PSO is very effortless. Contrast with the other developing calculations, it occupies the bigger optimization ability and it can be completed easily. It also assumes the real number code, and it is determined honestly by the solution. The number of the elements is equal to the constant of the solution. The selection of data mining approaches depends on the nature of the dataset if the dataset consist of the labeled features then the classification techniques can be suggested for best prediction. If the dataset is with unlabelled features then the clustering techniques are best suited for pattern recognition.

If the optimization of the results needs to be improvised means then bio inspirational based techniques are best suited. In still the diagnosis of disease suffers from high false alarm and detection rate is low in the future work we planned to propose a novel approach to reduce the false alarm rate in the situation of uncertainty handling.

7. REFERENCES

[1] Kumari M. and Godara S., “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction,” *International Journal of Computer Science and Technology (IJCST)* Vol. 2, Issue 2, June 2011.

[2] Imran Kurt, Mevlut Ture, A. Turhan KurumcComparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease, Elsevier Expert Systems with Applications, Volume 34, Issue 1, January 2008, Pages 366–374

[3] Carlos Ordonez.: Comparing association rules and decision trees for heart disease prediction, ACM, HICOM (2006)

[4] Sellappan Palaniappan et al, Intelligent heart disease prediction on system using data mining techniques.IJCSNS Vol 8 no 8(Aug 2008)

[5] Thangavel, K., Jaganathan, P.P. and Easmi, P.O (2006). *Data Mining Approach to Cervical Cancer Patients Analysis Using Clustering Technique*. Asian Journal of Information Technology

[6] Ekong, V.E., Onibere, E.A., Imianvan, A.A, Fuzzy Cluster Means System for the Diagnosis of Liver Diseases, IJCST Vol. 2, Issue 3, September 2011

[7] Wael A. Al Zoubi, Azuraliza Abu Bakar, Khairuddin Omar,” Scalable and Efficient Method for Mining Association Rules”, 2009 International Conference on Electrical Engineering and Informatics

[8] Wang, X. Y. and Garibaldi, J., 2005, "Simulated Annealing Fuzzy Clustering in Cancer Diagnosis", European Journal of Informatica

[9] Wang, X. Y., Whitwell, G. and Garibaldi, J., 2004, "The Application of a Simulated Annealing Fuzzy Clustering Algorithm for Cancer Diagnosis", Proceedings of the IEEE 4th International Conference on Intelligent System Design and Application, Hungary, 467-472

[10] Wang, X. Y., Garibaldi, J. and Ozen, T., 2003, "Application of The Fuzzy C-Means Clustering Method on the Ananlysis of non Pre-processed FTIR Data for Cancer Diagnosis", Proceedings of the 8th Australian and New Zealand Conference on Intelligent Information Systems, December 10-12, Australia, 233-238

[11] J. C.Bezdek, Cluster Validity with Fuzzy Sets. J. Cybernet, Vol. 3, No. 3, pp. 58-72, 1974.

[12] J. C. Bezdek, Mathematical Models for Systematics and Taxonomy. In Estabrook, G. (Ed.), Proceeding of 8th Internat Conference Numerical Taxonomy. Freeman, San Francisco, CA, pp. 143-166, 1975.

[13] X. L. Xie and G. Beni, " A validity measure for fuzzy clustering", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 13, pp. 841-847, 1991.

[14] N. R. Pal, J. C. Bezdek, "On cluster validity for the fuzzy c-means model", IEEE Trans. Fuzzy Systs., Vol. 3, pp. 370-379, 1995.

[15] Chowdhury SR, Chakrabarti D, Hiranmay S, Medical diagnosis using adaptive perceptive particle swarm optimization and its hardware realization using field programmable gate array, pubmed, 2009 Dec;33(6):447-65.

[16] Mona Nagy Elbedwehy, Hossam M. Zawbaa, Neveen I. Ghali, Aboul Ella Hassanien, Detection of Heart Disease using Binary Particle Swarm Optimization, IEEE Explore, Computer Science and Information Systems (FedCSIS), 2012

[17] K. Polat, S. Sahan, H. Kodaz, and S. Gunes, "Breast cancer and liver disorders classification using artificial immune recognition system (airs) with performance evaluation by fuzzy resource allocation mechanism," Expert Systems with Application, p. 172183, 2

[18] K. Polat and S. Gunes, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, p. 702710, 2007

[19] M. Ganji and M. Abadeh, "Using fuzzy ant colony optimization for diagnosis of diabetes disease," IEEE, 2010.

[20] T. Jayalakshmi and A. Santhakumaran, "A novel classification method for diagnosis of diabetes mellitus using artificial neural networks," in International Conference on Data Storage and Data Engineering, 2010.

[21] Amin Einipour , Global Journal of Health Science Vol. 3, No. 2; October 2011.