A Comparison of Hybrid Neural Network Based Breast Cancer Diagnosis Systems

Hsine-Jen Tsai[1(✉)], Hao-Chun Lu[1], Tung-Huan Wu[1], and Chiang-Sheng Lee[2]

[1] Fu Jen Catholic University, New Taipei, Taiwan, ROC
{tsai.fju, bach0809}@gmail.com,
discgto@yahoo.com.tw
[2] National Taiwan University of Science and Technology, Taipei, Taiwan, ROC
cslee@mail.ntust.edu.tw

Abstract. Breast cancer is the second leading cause of death among the women aged between 40 and 59 in the world. The diagnosis of such disease has been a challenging research problem. With the advancement of artificial intelligence in medical science, numerous AI based breast cancer diagnosis system have been proposed. Many researches combine different algorithms to develop hybrid systems to improve the diagnosis accuracy. In this study, we propose three artificial neural network based hybrid diagnosis systems respectively combining association rule, correlation and genetic algorithm. The effectiveness of these systems is examined on Wisconsin Breast Cancer Dataset. We then compare the accuracy of these three hybrid diagnosis systems. The results indicated that the neural network combining with association rule not only has excellent dimensionality reduction ability but also has the similar accurate prediction with correlation based neural network which has best accurate prediction rate among all three systems compared.

Keywords: Neural network · Association rule · Genetic algorithm · Medical artificial intelligence

1 Introduction

Breast cancer is the second leading cause of death among women in the United States according to the National Breast Cancer Foundation. The number of new cases of cancer in 2012 has reached around 14.1 million worldwide and 11.9 % (around 1.7 million) of these cases were diagnosed with breast cancer according to the WHO (World Health Organization). Breast cancer is a disease in which a malignant tumor forms in the tissues of the breast. A malignant tumor is a group of cancer cells that can grow into surrounding tissues in breast, but with early detection and treatment, most people continue to live a normal life. Early diagnosis is one of most significant steps in reducing the health and social complications of this disease. In the last decades, with increased emphasis towards cancer related research, new and innovative methods for

F.F.-H. Nah and C.-H. Tan (Eds.): HCIB 2015, LNCS 9191, pp. 633–639, 2015.
DOI: 10.1007/978-3-319-20895-4_59

early detection and treatment have been developed. Due to the use of electronic data capture and data management systems for both clinical care and biomedical research, the medical research has become toward quantitative research [9, 11, 14]. The abundance of data is strongly accelerating the trend. Data-driven study is becoming a common complement in medical diagnosis system. Many medical diagnosis systems use artificial neural networks (ANN) as a classification approach [2, 3, 5–8, 11, 12]. Artificial neural networks is a powerful tool which helps medical professionals to analyze, model and make sense of complex clinical data across a broad range of medical applications.

In this research, we proposed three artificial neural network (ANN) based hybrid diagnosis systems respectively combining association rule (AR), correlation and genetic algorithm (GA). The effectiveness of these systems is examined on Wisconsin Breast Cancer Dataset. The accuracy of these three hybrid diagnosis systems is compared. The main motivation behind this study is to use different approaches to minimize the number of features and then use the neural network to perform the prediction. By eliminating unnecessary features, we can save time and resource of computation during the prediction process.

In the next section we look at the literature review. Section 3 proposes three hybrid diagnosis systems which are artificial neural networks combining association rule, correlation and genetic algorithm respectively. Details of models and algorithm of these systems are described in this section as well. The experimental results are presented in Sect. 4. Finally, Sect. 5 provides the conclusions and future directions of research.

2 Literature Review

Breast cancer is the most common cancer in women both in the developed and less developed world. It is estimated that worldwide over 508,000 women died in 2011 due to breast cancer according to the WHO (World Health Organization) in 2013. Many research related to breast cancer have been reported and applied. They are prediction of breast cancer survivability [4, 13], reoccurrence rate and diagnosis of breast cancer [6, 10], etc. Many researchers have tried to use different methods to improve the accuracy of diagnosis system.

As the applications being developed in the data mining areas, researchers are still struggled with some challenges. Features selection is one of inevitable problems when there are significant amount of input features for a particular data mining applications. Limiting the number of input features has influence on the performance of data mining models in great part. Recently, many hybrid data mining systems have been put forward. Artificial neural network is one of the most common methods for prediction problems. Reference [5] proposed a hybrid model combing case-based reasoning and fuzzy decision tree and achieved 98.4 % forecasting accuracy for breast cancer. Reference [13] provided a diagnose model combing artificial neural network with genetic algorithm by processing patients' infrared thermal images to diagnose breast cancer.

3 Models and Algorithms

3.1 Database

The required data for this research was obtained from Wisconsin breast cancer database. They have been collected by Dr. William H. Wolberg at the University of Wisconsin-Madison Hospitals. There are 699 records in this database. Each record consists of nine features. These nine features detailed in Table 1 are graded on an interval scale from a normal state of 1–10, with 1 being the normal and 10 being the most abnormal state. 241 records out of 699 are malignant and 458 records are benign.

3.2 Models

Feature selection plays an important role in building a prediction model. By eliminating redundant input features that has no significant influence on the final outcome, we can build a prediction model with better efficiency and prediction accuracy. We propose three hybrid models that use different approaches, namely association rule (AR), genetic algorithm (GA) and correlation, to perform the feature selection task. Each model has two layers. First layer is the feature selector whose major task is to select significant features and lower the dimension of input vector. Second layer is the artificial neural network model to perform prediction. The general hybrid diagnosis system is shown in Fig. 1.

AR_Based ANN Model
Association rule is a method to discover relationship among items in large databases. A typical and well-known example of association rule is Market Basket Analysis [1]. That is, given a collection of items and a set of transactions, each transaction contains some number of items from given collection. An association algorithm can find rules such as 85 % of all the transactions that contain items A and B also contain items C and D.

Table 1. Descriptions of features in Wisconsin breast cancer database

Feature code	Feature description	Values of features	Mean	Standard deviation
A	Clump thickness	1–10	4.42	2.82
B	Uniformity of cell size	1–10	3.13	3.05
C	Uniformity of cell shape	1–10	3.20	2.97
D	Marginal adhesion	1–10	2.80	2.86
E	Single epithelial cell size	1–10	3.21	2.21
F	Bare nuclei	1–10	3.46	3.64
G	Bland chromatin	1–10	3.43	2.44
H	Normal nucleoli	1–10	2.87	3.05
I	Mitoses	1–10	1.59	1.71

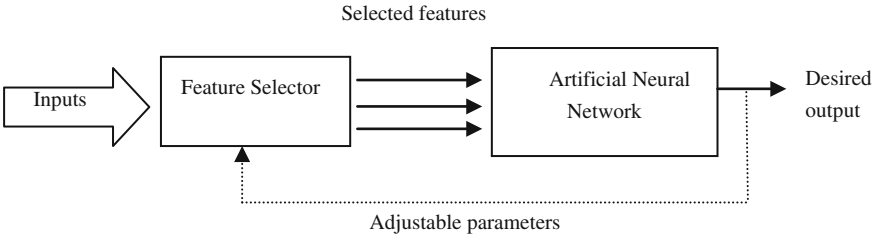

Fig. 1. The block diagram of the hybrid diagnosis system

Apriori algorithm [1] is used in the feature selector of AR_based ANN model. Most of the association rule algorithms are somewhat variations of this algorithm. The Apriori algorithm [1] is given as follows:

```
Apriori()
L1 ={large 1-itemsets}
k = 2
while Lk-1≠ϕ do
begin
Ck = apriori_gen(Lk-1)
for all transactions t in D do
begin
    Ct = subset(Ck, t)
    for all candiate c ∈ Ct do
    c.count = c.count+1
    end
    Lk ={ c ∈ Ck | c.count ≥ minsup }
    k = k + 1
end.
```

To run the Apriori algorithm, we use all input features and their all records to find some large itemset which has high confidence value and enough support value. For example, a large itemset [A, C, D, F] is obtained with 95 % of confidence and 80 % of support. Then feature A is selected as the representative of this itemset. The rest features in the itemset are redundant and eliminated. After several runs of such process with some sets of support and confidence, the input features of the second layer which is artificial neural network are obtained.

GN_Based ANN Model

Genetic algorithm is a common technique for optimization problems. In genetic algorithm, the population is associated with n chromosomes that represent candidate solution; each chromosome is an m-dimensional vector where m is the number of optimized parameters.

In our GA_based ANN model, the process of feature selection and prediction is stated as follows. At first, an input vector with a length of nine elements is created and feed into ANN model. Each element corresponds to the specific feature of the WBCD

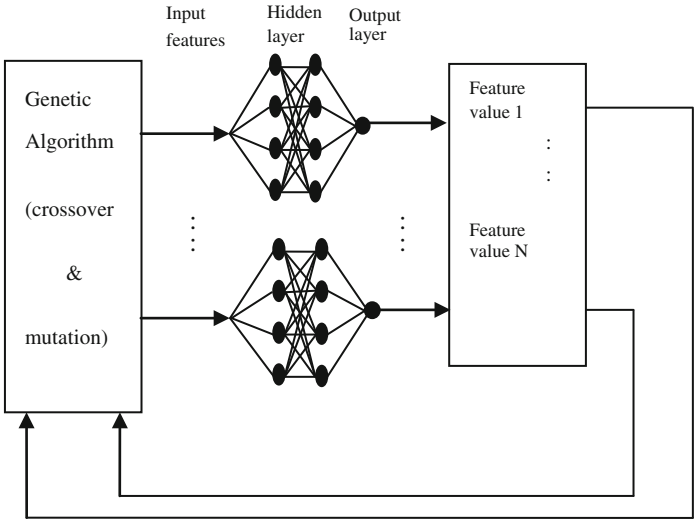

Fig. 2. A block diagram of GN_based NN diagnosis system

record. Output of ANN model is then feed into feature selector, genetic algorithm module in this case. Fitness value is calculated inside the feature selector. New generation of input features (chromosome) then are generated after crossover and mutation inside the feature selector. The process continues until stop criterion are satisfied. Figure 2 shows a block diagram of GN_based NN diagnosis system.

Correlation_Based ANN Model

Correlation is one the basic technique in statistic area. By discovering correlation between input features, redundant features can be located and eliminated. In this model, correlation is used as the feature selector. After features of WBCD records are feed into selector and calculated, the correlation of features are stored in the matrix. Redundant features are then eliminated according the threshold. Use the new feature set as the input of the second layer which is ANN model and perform training and testing.

4 Experimental Results

This experiment was conducted on the Wisconsin breast cancer database. In test stage, 10-fold cross validation method was applied. Experimental results are presented using confusion matrix to evaluate the accuracy of each approach. Table 2 shows the result using the ANN only without feature selection process. The result of these three hybrid

Table 2. The confusion matrix of ANN only without feature selection process

Confusion matrix		T	F
	P	97.5728	2.4271
	N	2.8169	97.1830

Table 3. The confusion matrix of AR_based ANN

Confusion matrix		T	F
	P	93.6893	6.3106
	N	3.9436	96.0563

Table 4. The confusion matrix of correlation_based ANN

Confusion matrix		T	F
	P	98.5436	1.4563
	N	2.5352	97.4647

Table 5. The confusion matrix of correlation_based ANN

Confusion matrix		T	F
	P	97.0873	2.9126
	N	3.9436	96.0563

Table 6. Accruacy rate comparison of ANN with three hybrid ANN model

	Feature	Accuracy
ANN	A ~ I	95.32 %
AR_based ANN(2)	B, F	94.10 %
Correlation_based NN(1) 90 %	A, B, D, E, F, G, H, I	95.88 %
GA_based ANN	B, C, G, H, I	94.73 %

models are shown in Tables 3, 4, and 5. A comparison of all four models is presented in Table 6.

The results show that correlation based neural network has the accurate prediction rate with 95.88 % which is the best among all three systems compared. With respect to dimensionality reduction, the result of AR_based ANN model is better than GA_based ANN and Correlation_based ANN.

5 Conclusion

A considerable amount of medical intelligence research has been conducted in the last decade. However, the researchers put more focus on diagnosis prediction systems. Many artificial intelligent techniques have been investigated to diagnose the breast cancer. This work has explored the accuracy of hybrid diagnosis models combining feature extraction with different classification techniques. Three artificial neural network based hybrid diagnosis systems respectively combining association rule, correlation and genetic algorithm. The effectiveness of these systems is examined on

Wisconsin Breast Cancer Dataset. The accuracy of these three hybrid diagnosis systems is compared.

The results indicated that the correlation based neural network has the best accurate prediction rate among all three systems compared. The artificial neural network combining with association rule not only has excellent dimensionality reduction ability but also has the similar accurate prediction with correlation_based ANN.

References

1. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases (1994)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Clarendon Press, Oxford (1996)
3. Choua, S.M., Leeb, T.S., Shaoc, Y.E., Chenb, I.F.: Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst. Appl. **27**, 133–142 (2004)
4. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2004)
5. Dybowski, R., Gant, V.: Clinical Applications of Artificial Neural Networks. Cambridge University Press, Cambridge (2007)
6. Er, O., Yumusak, N., Temurtas, F.: Chest disease diagnosis using artificial neural networks. Expert Syst. Appl. **37**(12), 7648–7655 (2010)
7. Fan, C.-Y., Chang, P.-C., Lin, J.-J., Hsieh, J.C.: A hybrid model combining case-based reasoning and fuzzy decision tree for medical data classification. Appl. Soft Comput. **11**, 632–644 (2011)
8. Floyd, C., Lo, J., Yun, A., Sullivan, D., Kornguth, P.: Prediction of breast cancer malignancy using an artificial neural network. Cancer **74**, 2944–2998 (1994)
9. Karabatak, M., Cevdet, M.: An expert system for detection of breast cancer based on association rules and neural network. Expert Syst. Appl. **36**, 3465–3469 (2009)
10. Mangasarian, O.L., Wolberg, W.H.: Cancer diagnosis via linear programming. SIAM News **23**(5), 1–18 (1990)
11. Owrang, O., Mehdi, M.: Association rules mining for breast cancer survivability prediction. http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf
12. Shortliffe, E.H.: Clinical Information Systems In the Era of Managed Care. Sea Island, GA (1993)
13. Zadeh, H.G., Haddadnia, J., Hashemian, M., Hassanpour, K.: Diagnosis of breast cancer using a combination of genetic algorithm and artificial neural network in medical infrared thermal imaging. Iran. J. Med. Phys. **9**, 265–274 (2012)
14. Vimla, L., Patel, A., Edward, H., Shortliffea, B., Stefanellic, M., Szolovits, D., Michael, R., Berthold, E., Bellazzic, R., Abu-Hanna, A.: The coming of age of artificial intelligence in medicine. Artif. Intell. Med. **46**, 5–17 (2009)