An Expert System for Detection of Breast Cancer Using Data Preprocessing and Bayesian Network

Amir Fallahi, Shahram Jafari*

School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
**Email of the corresponding author: Jafaris@shirazu.ac.ir*

Abstract

This paper presents an automatic system for detection of breast cancer using data preprocessing and Bayesian network. In this study, ReliefF algorithm is used for reducing the dimension of breast cancer database then a pre-processing is done on the data and ultimately Bayesian network classifier is used for classification. The system performance has been compared with model NN (neural network) and AR + NN (neural networks combined with association rules). The dimension of input feature space is reduced from nine to eight by using ReliefF. In test stage, 3-fold cross validation method was applied to the Wisconsin breast cancer database to evaluate the proposed system performance. The correct classification rate of proposed system is 98.1%.This research offered that the preprocessing is necessary on this data and combination of ReliefF and Bayesian network can be used to obtain fast automatic diagnostic systems for breast cancer.

Keywords: *Wisconsin breast cancer database, Weka, ReliefF.*

1. Introduction

In western countries, there is one out of eight women who suffers from breast cancer and its highest percentage in terms of age is between 40 to 50 years. Cancers are divided into two types, benign and malignant and breast cancer has no exception. If the cancer is benign under the conditions of early diagnosis, increasing age and treatment hope for patients will be very high. Studying patients and identifying all the factors influencing the disease, there are hopes to take useful steps toward this.

One the methods to identify breast cancer which is used more than the others, is Mammography. But it is frequently seen that different interpretation of radiologists about images is obtained from this way. Another method is Fine needle aspiration cytology (FNAC) and its accuracy is 90%. Therefore, it is better to discover another accurate method. Data classification process which is done on the past data is one of the most important issues in statistics and computer science. Diagnosis and medical issues are some of the applications of this study. One of the areas of the application of database analysis and pattern recognition are fault detection systems. Because of the facilities in modernity, lots of databases in medicine can be collected. These databases need specific techniques for analyzing, processing and effective use of them. Data mining and knowledge discovery in data base are methods to discover hidden concepts in massive data. Different kinds of methodologies are visualization, machine learning and statistical techniques which are recognized under the names of data classification, prediction and clustering. Here, some of these techniques in order to forecast and classification of breast cancer patients are mentioned. In 2003, a combination of neural network methods and decision trees to predict the return of this disease were used.

In 2004, neural network approach and intelligent multivariate adaptive regression for classifying were used. In 2007, the harmonic separation techniques to predict the disease were used. Again in 2007 a combination of fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis was used. AR+NN method in 2009 for the issue of breast cancer diagnosis was used which included two-step approach. In this the work, AR (association rules) for reducing database-related aspects of cancer data and NN (neural network) for classifying the data were used. The accuracy of this method was 97.4%. This paper represents a method to diagnose breast cancer also it compares its results with AR+NN method.

In this model, at first, the number of database features is decreased by feature selection algorithm ReliefF, and then in the next stage, while deleting missing data, according to unbalanced data to classes, the data is balanced by SMOTE algorithm. In the final stage, by using Bayesian Network, data classification has been done and the accuracy is estimated which is 98.1%. For testing methods, a three stage valid test is used. This is because the results are compared with AR+NN method which were presented in 2009. All of the algorithms and tests in this paper have been done on the data by Weka tools.

2. Characteristics of Wisconsin Breast Cancer Database

Considering the importance of breast cancer and that it is prevalent among women, much research has been done in this area and scientists are making great progress in understanding how certain changes in DNA can cause normal breast cells to become cancerous. In this study, the Wisconsin breast cancer database was used and analyzed. They have been collected by Dr. William H. Wolberg (1989-1991) at the University of Wisconsin–Madison Hospital.

There are 699 records in this database. Each record in the database has nine attributes. The nine attributes are detailed in Table 1. Values of these fields are between 1 to10 of that 10 shows most unusual situation. This database includes 241 malignant and 458 benign records and database is imbalance. One feature of this database is that 16 records from the database, due to lack of data entry for one of their features, are incomplete. Due to the characteristics of the database and the importance of unbalancing and missing data for more accurate classification and data analysis, such a database was selected.

Table 1 Wisconsin breast cancer database contains 241 malignant and 458 benign

Attribute number	Attribute description	values	Missing rate
1	Clump thickness	1-10	0
2	Uniformity of cell size	1-10	0
3	Uniformity of cell shape	1-10	0
4	Marginal adhesion	1-10	0
5	Single epithelial cell size	1-10	0
6	Bare nuclei	1-10	2
7	Bland chromatin	1-10	0
8	Normal nucleoli	1-10	0
9	Mitoses	1-10	0

3. Feature Reduction

In this section in order to obtain more accurate classification the numbers of database features are decreased. In feature selection, the goal is to find a subset of significant attributes capable to correctly predict unseen data and to reduce both human measurement errors and the cost of the activity of data extraction. Ranking of features is possible since a large number of feature evaluation measures are available. We will rank the attributes using a ReliefF[2] evaluator. The key idea of Relief is to estimate the quality of features according to how well their values distinguish between the instances of the same and different classes that are close to each other.

ReliefFAttributeEval is used in conjunction with the Ranker search method of Weka to generate a ranked list; this is based on Relief algorithm. ReliefFAttributeEval is instance-based and it samples instances randomly and checks nearby instances of the same and different classes. This evaluator gives a numeric value to each feature that indicates importance and the value of the feature. Accordingly, a threshold value for evaluator can be set for determination of what extent (less than or equal a number) is not significant. By doing this, unimportant features can be decreased. Therefore, ReliefFAttributeEval evaluator on breast cancer database adjusts these parameters and has been done by a Weka tool.

Parameter	**value**	**Description**
-M	-1	number of instances where -1 means all instances of our Dataset
-D	1	Seed for randomly sampling instances
-K	10	number of neighbors
Threshold	**-1.7976**	

At the end of this stage and after feature reduction by ReliefF, feature “I” was selected for being reduced.

4. Missing Data Handling and Data Balancing

As previously noted two important features of this database are missing Data and data unbalancing. In this section, compensation of these defects will be tried to be able to provide more accurate classification results. First, different methods to address missing data that are used much, like mode, average or minimum value of feature in which the missing data occurred to fill the data were used. But the best was when the instance containing missing data were excluded.

So these instances that numbers of them were 16 instances were removed from the Wisconsin database. Now it is the turn for data unbalancing correction that for this problem SMOTE should be used [3] a method that is considered as a over-sampling technique.

In SMOTE approach minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. Their implementation currently uses five nearest neighbors.

For instance, if the amount of over-sampling needed is 200%, only two neighbors from the five nearest neighbors are chosen and one sample is generated in the direction of each. Synthetic samples are generated in the following approach: take the difference between the sample under consideration and its nearest neighbor; multiply this difference by a random number between 0 and 1, and add it to the feature vector under consideration.

This causes the selection of a random point along the line segment between two specific features. This approach effectively forces the decision region of the minority class to become more general. SMOTE algorithm in Weka was used; the only important parameter in this algorithm was the number of nearest neighbors that it was set number 5 for that. With the run of this algorithm the numbers of samples in two classes were balanced and thus the numbers of samples in the class of benign were 444 and samples in the malignant class were changed to 478. While data balancing after removing 16 instances that contain missing data was performed. At the end of this stage the database using an attribute that contains missing data, was arranged.

5. Bayesian Network

A Bayesian network, belief network or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. In this model nodes represent random variables.

They may be observable quantities, latent variables, unknown parameters or hypotheses. If there is an arc from node A to another node B, A is called a parent of B, and B is a child of A. The set of parent nodes of a node xi is denoted by parents (xi). A directed acyclic graph is a BN relative to a set of variables if the joint probability distribution of the node variables can be written as the product of the local distributions of each node and its parents as:

$$P(x_1, \ldots x_n) = \prod_{i=1}^{n} P(x_i \,|\, \mathrm{parents}(x_i)) \qquad \text{Equation}(1)$$

The aim of supervised classification is to classify instances i given by certain characteristics xi={xi1,…,xin} into r class labels, ci , i=1,…r. xil denotes the value of variable xl observed in instance i. The main principle of a Bayesian classifier is the application of Bayes' theorem. Bayes' theorem, Eq. (2), calculates the posterior probability P (cj|xi) from the conditional probabilities P(xj|ck) and the prior probabilities P(ck) as:

$$P(c_j|x_i) = \frac{P(x_i\,|c_j)P(c_j)}{\sum_k P(x_i\,|c_k)P(c_k)} \qquad \text{Equation}(2)$$

The posterior probability P (cj|xi) is the probability that a sample with characteristics xi belongs to class cj . The prior probability P(cj) is the probability that a sample belongs to class cj given no information on its characteristic values. The probabilities of Eq. (2) can be estimated from the expert or from a training set required to build the classifier, where each instance i is given by (xi,cj) .Bayes' theorem is used to predict the class and classify each unseen instance: a new instance or example j, only characterized with the values xj of the predictor variables, is given a class label according to the class that has the maximum posterior probability.

A useful property of the Bayesian classifier is that it is optimum in the sense that the expected rate of misclassifications is reduced to a minimum. Thus the data after preprocessing by Bayesian Net classifier were classified. Like the rest of the steps work, job classification with the following settings in the Weka tool was done.

```
Estimator  :SimpleEstimator  -A 5
Search algorithm : K2  -P 1 –S BAYES
```

6. Results

In this section we compare our results with AR+NN and NN models. Our results are produced from a 3-fold cross validation method and average values were calculated. The performance comparison and correct classification rates are shown in Table 2.

6-1 Accuracy obtained for a Bayesian network approach in the fourth row shows that this method, even without removal of any feature of the data and perform any preprocessing on the data, is better than a neural network approach for breast cancer diagnosis.

6-2 In the next section, it can be seen that the removal of a feature by ReliefF algorithm and using the Bayesian network classifier, although very low but still is better than the combination of neural network algorithm and AR works.

Table 2 Results of Comparison within Classifiers

The classifier	**Correct classification rate(%)**
NN(9,11,1)	95.2
AR1+NN(8,11,1)	97.4
AR2+NN(4,11,1)	95.6
(Our method) Bayesian Net	97.28
(Our method) Bayesian Net +ReliefF	97.42
(Our method) Bayesian Net+data balancing +remove missing data	97.83
(Our method) Bayesian Net+ReliefF+remove missing data+data balancing+sorting data upon feature F	98.15

6-3 The next results show that the data preprocessing for classification has a significant impact on accuracy of classifiers. Best state is when feature “I” is removed from the database by using ReliefF algorithm, then the missing data are resolved and the data are balanced by using SMOTE algorithm (an over-sampling technique).

After doing these things, the data are sorted based on “F” feature. It is experimentally shown that can lead to better results for the data classification that this is related to the nature of the classifiers. Finally, the prepared data are classified by using Bayesian network classifier. The best accuracy achieved for this method is 98.15%.

7. Conclusion

In this study, an automated diagnosis system to detect breast cancer based on data preprocessing and Bayesian networks is presented. This study shows that Bayesian networks in the diagnosis of the disease according to the nature of the algorithms work very well. ReliefF algorithm is also used to reduce the number of the database features that shows it can produce useful results in combination with Bayesian network classifier.

Also, we tried to show that the classifiers can reach to a certain extent of accuracy for this database. Thus for obtaining more accuracy we should do enough preprocessing on the data, resolve missing data and data unbalancing phenomenon for the data. Finally, we showed that we can obtain significant accuracy for diagnosing breast cancer disease by using 3-fold cross validation test method.

References

[1] M. Karabatak, M. Cevdet, "An expert system for detection of breast cancer based on association rules and neural network", Expert Systems with Applications, 36, (2009), pp. 3465–3469.

[2] K. Kira and L. Rendell, "A pratical approach to feature selection", In D.Sleeman and P. Edwards, editors, Proceedings of the Ninth International Workshop on Machine (1992).

[3] Nitesh V. Chawla, "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, Vol.16, (2002), pp. 321-357.

[4] Friedman, N., Geiger, D., & Goldszmit, M., "Bayesian network classifiers", Machine Learning, 29, (1997), pp. 131–161.

[5] Yetian Chen, "Learning Classifiers from Imbalanced, Only Positive and Unlabeled Data Sets", CS573 Project, (2009).

[6] Marko Robnik-Sikonja, Igor Kononenko, "An adaptation of Relief for attribute estimation in regression", In: Fourteenth International Conference on Machine Learning, (1997), pp. 296-304.

[7] Übeyli, E. D., "Implementing automated diagnostic systems for breast cancer", (2007).

[8] M. Correa, C. Bielza , J. Pamies-Teixeira, "Comparison of Bayesian networks and artificial neural networks for quality detection in a machining process", Expert Systems with Applications, 36, (2009), pp. 7270–7279.

[9] Aragonés, M. J., Ruiz, A. G., Jiménez, R., Pérez, M., & Conejo, E. A., "A combined neural network and decision trees model for prognosis of breast cancer relapse", Artificial Intelligence in Medicine, 27, (2003), pp. 45–63.

[10] D. Heckerman, D. Geiger, D. M. Chickering, "Learning Bayesian networks: the combination of knowledge and statistical data", Machine Learning, 20, (1995), pp. 197-243.

[11] I.H. Witten, E. Frank Morgan Kaufmann, "Data mining: Practical machine learning tools and techniques with Java implementations", (2000).

[12] R.R. Bouckaert, "Bayesian Belief Networks: from Construction to Inference", Ph.D. thesis, University of Utrecht, (1995).

[13] G. Cooper, E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data". Machine Learning, 9, (1992), pp. 309-347.