International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org ||Volume 5 Issue 12|| December 2016 || PP. 62-65

Classification of Breast Cancer Diseases using Data Mining Techniques

Akhilesh Kumar Shrivas[1], Ankur Singh[2]
Dept. of IT, Dr. C. V. Raman University, Kota , Bilaspur(C.G.)

Abstract: *Medical data mining has great deal for exploring new knowledge from large amount of data. Classification is one of the important data mining techniques for classification of data. In this research work, we have used various data mining based classification techniques for classification of cancer diseases patient or not. We applied the Breast Cancer-Wisconsin (Original) data set into different data mining techniques and compared the accuracy of models with two different data partitions. BayesNet achieved highest accuracy as 97.13% in case of 10-fold data partitions. We have also applied the info gain feature selection technique on BayesNet and Support Vector Machine (SVM) and achieved best accuracy 97.28% accuracy with BayesNet in case of 6 feature subset.*

Keywords: *Classification, Data Mining, Decision Tree, Cross Validation.*

I. Introduction

In medical science, patient data is increasing every day due to huge number of patients. Data mining plays very important role to extract useful information from large amount of data. The main objective of this research work is to analysis and classifies the cancer diseases data using different data mining techniques. There are various authors have worked in the field of classification of cancer diseases. E. I. Papageorgiou et al. [1] has proposed FCM (Fuzzy cognitive method) techniques for classification of breast cancer. The accuracy of the proposed model given 95% accuracy. B. Zheng, et al. [2] have suggested K-SVM model for classification of breast cancer data. M. Kanchana et al. [3] have presented a probabilistic neural network for the classification of the breast cancer based on discrete wavelet transform using digital mammogram images. The proposed Computer Aided Diagnosis (CAD) system is tested using MIAS (Mammography Image Analysis Society) database and achieved an accuracy of 92.3%. M. Karabatak et al. [4] have suggested an automatic diagnosis system for detecting breast cancer based on association rules (AR) and neural network (NN) and achieved 95.6% of accuracy. P. Thangaraju et al. [10] have suggested various classification techniques like J48, Bayes Net, Naïve byaes, decision table, multilayer perceptron, REP Tree etc. for classification of breast cancer. They recommended decision table as best model. Similarly, there are various authors have used different data mining techniques, soft computing techniques and statistical techniques for classification of breast cancer disease patient.

II. Methodology

Data mining techniques play important role to development the model. Classification is one of the important applications of data mining for classification of data. In this research work, we have used different classification techniques for classification of cancer diseases.

(i) C4.5

C4.5 [5] is an extension of ID3 that handle the unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. In building a decision tree, we can deal with training sets that have records with unknown attributes values by evaluating the gain, or the gain ratio, for an attribute values are available. We can classify the records that have unknown attribute value by estimating the probability of the various possible results. C4.5 produces tree with variable branches per node. When a discrete variable is chosen as the splitting attribute in C4.5, there will be one branch for each value of the attribute.

(ii) Classification and Regression Technique (CART)

CART (Classification and Regression Tree) [5] is one of the popular data mining techniques of building decision tree. It builds a binary decision tree by splitting the record at each node, according to a function of a single attribute. CART uses the gini index for determining the best split. The initial split produces the nodes, each of which we now attempt to split in the same manner as the root node. Once again, we examine the entire input field to find the candidate splitters. If no split can be found then significantly decreases the diversity of a given node, we label it as a leaf node. Eventually, only leaf nodes remain and we have grown the

full decision tree. The full tree may generally not be the tree that does the best job of classifying a new set of records, because of overfitting.

(iii) Bayesian Net

Bayesian classifiers [6] are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem. Classification algorithms have found a simple Bayesian classifier known as the Naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers. Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

(iv) Support Vector Machine (SVM)

Support vector machines (SVMs) [7] are supervised learning methods that generate input-output mapping functions from a set of labelled training data. The mapping function can be either a classification function (used to categorize the input data) or a regression function (used to estimation of the desired output). For classification, nonlinear kernel functions are often used to transform the input data to a high dimensional feature space in which the input data becomes more separable (i.e., linearly separable) compared to the original input space. SVMs belong to a family of generalized linear models which achieves a classification or regression decision based on the value of the linear combination of features. They are also said to belong to "kernel methods". In addition to its solid mathematical foundation in statistical learning theory, SVMs have demonstrated highly competitive performance in numerous real-world applications, such as medical diagnosis, bioinformatics, face recognition, image processing and text mining, which has established SVMs as one of the most popular, state-of-the-art tools for knowledge discovery and data mining.

(v) Multilayer Perceptron (MLP)

MLP [5] is a development from the simple perceptron in which extra hidden layers (layers additional to the input and output layers, not connected externally) are added. More than on hidden layer can be used. The network topology is constrained to be feed forward, i.e., loop-free. Generally, connections are allowed from the input layer to the first (and possible only) hidden layer, from the first hidden layer to the second and so on, until the last hidden layer to the output layer. The presence of these layers allows an ANN to approximate a variety of non-linear functions. The actual construction of network, as well as the determination of the number of hidden layers and determination of the overall number of units, is sometimes of a trial-and-error process, determined by the nature of the problem at hand. The transfer function generally a sigmoid function.

(vi) **Radial Basis Function (RBF)**

The Radical Basis Function (RBF) network [8] is popular several times. The popularity of this network arises from the two basic facts. The first one is that unlike most supervised learning neural network algorithms, it is able to find global optimum. For comparison, using a feed forward neural network with the back propagation learning rule usually finds only the local optimum. The second fact is that training time for RBF network is short compared with the other neural network, most notably when using the back propagation rule for adjustment of the weights. In addition, the topology of the RBF network is very simple to set up, and requires no guessing as with back propagation.

III. Data Set and Performance Measures

The Cancer-Wisconsin (Original) data is collected from UCI repository [9]. The data set contains 699 samples in which 458 samples are benign and 241 samples are malignant. The number of features in data set is 10 and 1 class. The class level is binary class as benign and malignant. The data set contains also missing value.

The performance measures are very important factors to analyze the robustness of model. The confusion matrix is very important factor for calculating various performance measures like accuracy, sensitivity and specificity. Confusion matrix [6] consist true positive (TP), true negative (TN), false positive (FP) and false negative (FN). Table 1 shows that confusion matrix and table 2 shows that various performance measures.

Table 1: Confusion matrix

Actual Vs. Predicted	**Positive**	**Negative**
Positive	TP	FN
Negative	FP	TN

Table 2: Performance measures

Measures	**Equation**
Accuracy	(TP+TN)/N
Sensitivity	TP/ (TP +FN)
Specificity	TN/ (TN +FP)

IV. Results And Discussion

This research work done in WEKA data mining software [11] with window environment. In this research work, used various data mining based classification techniques like C4.5, CART, SVM and BayesNet .We have applied Breast Cancer-Wisconsin (Original) data set on different classification techniques with 10-fold cross validation and 80-20% training-testing partitions. Table3 show that various performance measures like accuracy, sensitivity and specificity of different models with 10-fold cross validation and 80-20% data partition. We have achieved highest accuracy in bayes net as 97.13% and 96.42% with 10-fold cross validation and 80-20% partition respectively. Fig. 1 shows that accuracy of different models with 10-fold cross validation and 80-20% training-testing data partitions. The other performance measures like sensitivity and specificity is better in case of byesNet. SVM gives similar accuracy as bayes net in most of case with both the partition.

We have also applied the Info gain feature selection technique on Breast Cancer-Wisconsin (Original) data set and rank the features of data set. The rank of features of data set from high to low important are 2,3,6,7,5,8,1,4 and 9. We have removed the feature one by one form data set and applied to the bayesNet and SVM and calculate the accuracy with reduced number of features. Table 4 shows that accuracy of SVM and Bayes Net with different feature subsets. We have achieved 96.85% and 96.42% of accuracy in 7 and 6 features with 10-fold cross validation and 80-20% data partition respectively in case of SVM. Similarly, achieved 97.28% and 96.42% of accuracy in 6 and 4 features with 10-fold cross validation and 80-20% data partition respectively in case of Bayes Net. We can recommend both the models are better for classification of breast cancer.

Table 3: Performance measures of models

Model	10-fold cross validation			80-20% partitions		
	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity
C4.5	94.563	95.633	92.531	92.857	91.111	96
CART	94.849	95.414	93.775	95.00	95.555	94
SVM	**96.995**	97.379	96.265	**96.428**	96.666	96
Bayes Net	**97.138**	96.506	98.340	**96.428**	96.666	96
MLP	95.279	96.069	93.775	95.714	95.555	96
RBF	95.851	95.414	96.680	95.714	95.555	96

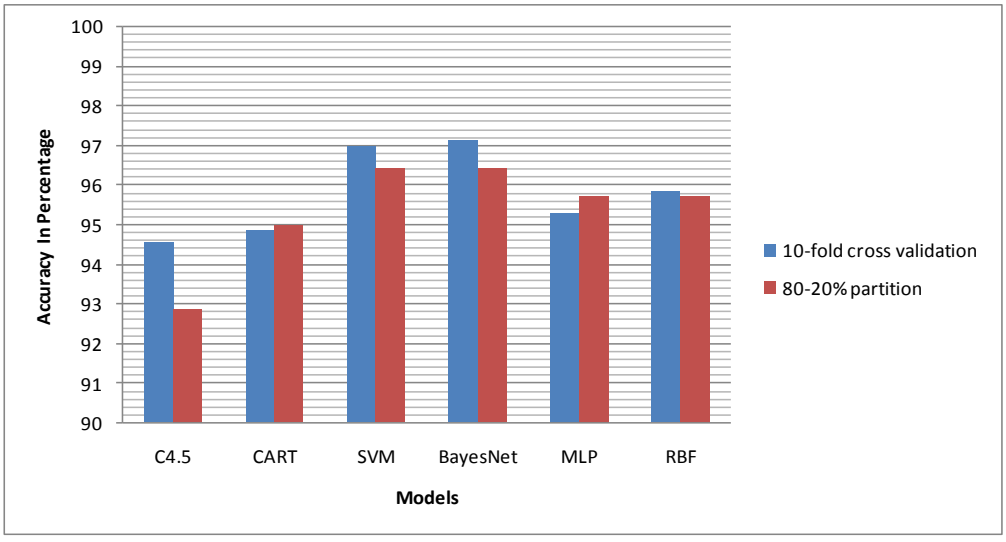

Fig.1: Accuracy of models with two different data partitions

Table 4: Accuracy of best model with Info Gain feature selection technique

Number of Feature	10-fold cross validation		80-20% partitions	
	SVM	**BN**	**SVM**	**BN**
8	96.709	97.424	96.428	96.428
7	**96.852**	97.281	96.428	96.428
6	96.137	**97.281**	**96.428**	96.428
5	95.565	96.423	95.714	96.428
4	95.565	96.280	96.428	**96.428**
3	95.422	95.422	96.428	95.714
2	93.705	94.277	95.714	95

V. Conclusion And Future Work

To extract the useful information from large amount of data is very important task for every organization and institution. Classification is one of the important data mining applications for classification of data. The main focus of this research work is to develop the robust model for identifying diseases and classifying diseases with high accuracy. Feature selection is also beneficial for computationally increase the performance of model. We have recommended SVM and Bayes Net with Gain ratio feature selection technique which gives satisfactory accuracy for classification of diseases. In future, we will develop hybrid and ensemble model to improve the classification accuracy. We will also use other feature selection techniques like gain ratio, genetic algorithm, particle swarm optimization etc. to computationally increase the performance of model.

References

[1]. E. I. Papageorgiou, J. Subramanian, A. Karmegam and N. Papandrianos, A risk management model for familial breast cancer: A new application using fuzzy cognitive map method, Elsevier, Computer methods and program in biomedicine, DOI:10.1016/j.cmpb.2015.07.003 ,2015, 1-13.

[2]. B. Zheng, S. W. Yoon and S. S. Lam, Breast cancer diagnosis based on feature extraction using a hybrid a k-means and support vector machine algorithm, Expert System with application ,41, 2014,1476-1482 .

[3]. M. Kanchana and P. Varalaxmi, Breast cancer diagnosis using Wavelet based threshold Method", East Journal of Scientific Research, 23, 2015, 1030-1034 .

[4]. M. Karabatak and M. Cavedet Ince, An Expert System for detection of Breast cancer based on association rules and Neural network, Expert System with applications, 36, 2009, 3465-3469.

[5]. A. K. Pujari, Data Mining Techniques,. Universities Press (India) Private Limited, 4th ed., ISBN: 81-7371-380-4, 2001.

[6]. J. Han, and M. Kamber , Data Mining Concepts and Techniques, Morgan Kaufmann, San Francisco, 2nd ed., ISBN: 13: 978-1-55860-901-3,2006.

[7]. D. L. Olson and D. Delen, Advanced Data Mining Techniques, USA, Springer Publishing: ISBN: 978-3-540-76916-3,2008.

[8]. K. J Cios., W. Pedrycz, and R. W. Swiniarski , Data Mining Methods for Knowledge Discovery. Kluwer Academic Publishers, 3rd ed., ISBN: 0-7923-8252-8,1998.

[9]. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/datasets.html].

[10]. P. Thangaraju and R. Mehala, Novel Classification based approaches over Cancer Diseases, International Journal of Advanced Research in Computer and Communication Engineering,4,2015,294-297.

[11]. Web source: http:// www.cs.waikato.ac.nz/~ml/weka/ last accessed on Aug. 2016.