Journal of Computer Science 6 (5): 576-584, 2010
ISSN 1549-3636

Predicting the Severity of Breast Masses with Ensemble of Bayesian Classifiers

Alaa M. Elsayad
Department of Computers and Systems, Electronics Research Institute,
12622 Bohoth St., Dokki, Geza, Egypt

Abstract: Problem statement: This study evaluated two different Bayesian classifiers; tree augmented Naive Bayes and Markov blanket estimation networks in order to build an ensemble model for prediction the severity of breast masses. The objective of the proposed algorithm was to help physicians in their decisions to perform a breast biopsy on a suspicious lesion seen in a mammogram image or to perform a short term follow-up examination instead. While, mammography is the most effective and available tool for breast cancer screening, mammograms do not detect all breast cancers. Also, a small portion of mammograms show that a cancer could probably be present when it is not (called a false-positive result). **Approach:** Apply ensemble of Bayesian classifiers to predict the severity of breast masses. Bayesian classifiers had been selected as they were able to produce probability estimates rather than predictions. These estimated allow predictions to be ranked and their expected costs to be minimized. The proposed ensemble used the confidence scores where the highest confidence wins to combine the predictions of individual classifiers. **Results:** The prediction accuracies of Bayesian ensemble was benchmarked against the well-known multilayer perceptron neural network and the ensemble had achieved a remarkable performance with 91.83% accuracy on training subset and 90.63% of test one and outperformed the neural network model. **Conclusion:** Experimental results showed that the Bayesian classifiers are competitive techniques in the problem of prediction the severity of breast masses.

Key words: Breast cancer, data mining, prediction, Bayesian network, neural network

INTRODUCTION

Importance of machine learning in breast cancer diagnosis: Breast cancer is a very common and serious cancer for women. It is the second largest cause of cancer deaths among women. Mammography is one of the most used methods to detect this kind of cancer (Choua *et al*., 2004; Singh and Al-Mansoori, 2000). The value of mammography is that it can identify breast abnormalities with 85-90% accuracy. In literature, radiologists show considerable variation in interpreting a mammography. In such cases, Fine Needle Aspiration Cytology (FNAC) is adopted. But, the average correct identification rate of FNAC is only 90% (Elmore *et al*., 1994). It is necessary to develop better identification method to recognize the breast cancer. Computer aided diagnosis can help to reduce the number of false positives and therefore reduce the number of unnecessary biopsies. Statistical techniques and artificial intelligence methods have been successfully used to predict the breast cancer by several researchers (Kovalerchuck *et al*., 1997; Pendharkar *et al*., 1999). The objective of these identification techniques is to assign a patient to either a benign group that does not have breast cancer or a malignant group who has strong evidence of having breast cancer. These diagnostic problems are widely discussed as classification problems (Han and Kamber, 2006; Larose, 2006; Nisbet *et al*., 2009; Johnson and Wichern, 2002). However, there is a strong argument to treat such problems as tasks of learning class probability estimates from data.

Probability estimation classifiers: A probability estimation classifier estimates the conditional probability distribution of the values of the class attribute given the values of the predictive attributes. Such classification models which represent conditional distribution will be concise and easy to comprehend. They include Naive Bayes, logistic regression, decision tree and Bayesian network. Naive Bayes and logistic regression models can only represent simple distributions, whereas decision tree models can represent arbitrary distributions, but they fragment the training dataset into smaller and smaller pieces, which unavoidably yield less reliable probability estimates.

Bayesian Network (BN) is the best-known classifier that able to provide the probability distributions concisely and comprehensibly (Witten and Frank, 2005). BN is a probabilistic model that consists of dependency structure and local probability. BN is drawn as a network of nodes, one for each attribute, connected by directed edges in such a way that there are no cycles; a directed acyclic graph. The major advantage of BN is the ability to represent and hence understand knowledge. Recently, there is increasing attention regarding the application of BN in medical contexts (Linda *et al*., 2008). BN classifiers have been evaluated as potential tools for the diagnosis of breast cancer using two real-world databases in (Cruz-Ramirez *et al*., 2007; 2009). In this study, two different implementations of BN have been investigated for the prediction of severity of breast masses; Tree Augmented Naive Bayes (TAN) and Markov Blanket Estimation (MBE) learning algorithms. Both algorithms use Naive Bayes classifier as a starting point for the learning procedure. The class attribute is the single parent of each node of a Naive Bayes network: TAN considers adding a second parent to each node. While MBE ensures that every attribute in the data is in the Markov blanket of the node that represents the class attribute. This study proposes an ensemble of three BN networks to efficiently predict the severity of breast masses. Ensemble based methods enable an increase in generalization performance by combining individual BN networks train on the same dataset. The idea is to employ multiple models to do better than a single on often even the retrospective best of the individual models. The performances of these BNs and their ensemble are benchmarked against the Multilayer Perceptron Neural Network (MLPNN). The mammographic mass dataset contains BI-RADS assessment, attributes, the patient's age and type of severity (Elter *et al*., 2007; American College of Radiology, 1998). Each mass sample has to be classified into a benign or a malignant group.

MATERIALS AND METHODS

About the dataset: A radiologist is a physician who analyzes the radiograph to decide if there is a tumor or just normal tissue and whither the existing tumor is malignant (cancerous) or benign (gentle). Due to the variations in mammography interpretations, the problem is gotten ahead to the pathologist. A pathologist is a physician who analyzes cells and tissues under a microscope to determine whether they are malignant or benign. The pathologist's report helps characterize specimens taken during biopsy or other surgical procedures and helps determine treatment. To determine a tumor's histologic grade, a sample of breast cells must be taken from a breast biopsy, lumpectomy or mastectomy. The purpose of this study is to increase the ability of physicians to determine the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age. The objective is to reduce the high number of unnecessary breast biopsies. The six BI-RADS reporting categories represent gradations of the likelihood that a cancer exists, from lowest to highest probability. The mammographic mass dataset used here has been collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006 (Elter *et al*., 2007). BI-RADS stands for the Breast Imaging and Reporting Data System and was developed by the American College of Radiology (ACR), in collaboration with multiple other organizations in 1991 to present answers concern about ambiguous mammography reports with indecisive conclusions from radiologists (American College of Radiology, 1998). The data set is available by http access of the University of California at Irvine (UCI) machine learning repository (Asuncion and Newman, 2007; Blake and Merz, 1998). Table 1 shows the mammographic mass dataset which contains the BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity attribute) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms. The values of ordinal attribute represent categories with some intrinsic ranking while they nominal attribute represent categories with no intrinsic ranking in nominal type.

Table 1: Attributes of mammographic mass dataset

	Type	Values and labels		No. of missing
Attribute		Value	Label	values
BI-RADS assessment (non-predictive)	Ordinal	0	Assessment incomplete	2
		1	Negative	
		2	Benign findings	
		3	Probably benign	
		4	Suspicious abnormality	
		5	Highly suggestive of malignancy	
Ages	Integer		Patient's age in years	5
Mass shape	Nominal	1	Round	31
		2	Oval	
		3	Lobular	
		4	Irregular	
Mass margin	Nominal	1	Circumscribed	48
		2	Microlobulated	
		3	Obscured	
		4	Ill-defined	
		5	Speculated	
Mass density	Ordinal	1	High	76
		2	Iso	
		3	Low	
		4	Fat-containing	
Severity (target class)	Binominal	0	Benign	
		1	Malignant	

Bayesian networks: Several classification algorithms have been developed in the field of data mining information systems. Some of these algorithms are able to produce probability estimates rather than predictions. That is for each class label, they estimate the probability that a given sample belongs to that class. Probability estimates are often more useful than plain predictions. They allow predictions to be ranked and their expected costs to be minimized. BNs among other models are ones of these classification approaches. The benefits of BNs are that they present well-founded methods to represent any arbitrary probability class distributions concisely and comprehensibly in a graphical manner. BN model is drawn as a network of nodes, one for each attribute, connected by directed edges in such a way that there are no cycles. In other words, a BN is a directed acyclic graph consisting of (Cheng *et al*., 1998):

- Nodes (or small circles), that stand for random attributes; edges (or arrows), which represent probabilistic relationships among these attributes
- For each node, there exists a local probability distribution attached to it that depends on the state of its parents

BN consists of a qualitative part (structural model) that presents a visual representation of the interactions among attributes and a quantitative part (set of local probability distributions), which provides probabilistic inference and numerically measures the effect of attributes on each other. The qualitative and quantitative parts mutually determine a unique joint probability distribution over the attributes in a specific problem (Cooper, 1999). The main idea within the structure of BN is that of independence. This idea refers to the case where the instantiation of a specific attribute leaves other two attributes independent of each other. BN model allows the representation of a joint probability distribution in a compact and economical way by making extensive use of conditional independence, as shown in Eq. 1:

$$P(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} P(X_i \mid P_a(X_i)) \quad (1)$$

Where:

$P(X_1, X_2, ..., X_n)$ = The joint probabilities of attributes $X_1, X_2, ..., X_n$

$P_a(X_i)$ = The set of parent nodes of X_i; i.e., nodes with edges pointing to X_i

$P(X_i|P_a(X_i)|))$ = The conditional probability of X_i given its parents

Equation 1 shows how to pick up a joint probability from a product of local conditional probability distributions; such representation may be used to solve classification problems (Linda *et al*., 2008; Cruz-Ramirez *et al*., 2007; 2009; Cheng *et al*., 1998). The learning algorithm for BN has to contain two components:

- A function for evaluating a given network (goodness of fit measure)
- A method for searching through the space of possible networks

Normally, the learning algorithm starts with a given ranking of the attributes (i.e., nodes). Then it processes each node in turn and greedily adds edges from previously processed nodes to the current one. In each step it selects the edge that maximizes the network's score. If there is no additional enhancement, attention goes to the next node. The Naive Bayes (NB) classifier is one of the most effective methods to build BNs (Friedman *et al*., 1997). However, it works well only for simple distributions. Usually, NB network is used as a starting point for the search. In this study, two learning algorithms have been used to build the BN classifiers starting NB network; Tree Augmented Naive Bayes (TAN) and Markov Blanket Estimation (MBE) learning algorithms.

Markov Blanket Estimation (MBE): MBE is a learning algorithm to create BN model by identifying the conditional independence relationships among the attributes. This algorithm ensures that every attribute in the dataset is in the Markov blanket of the node that represents the class attribute (Witten and Frank, 2005). A node's Markov blanket includes all its parents, children and children's parents. Hence, if a node is absent from the class attribute's Markov blanket, its value is completely irrelevant to the classification. Using statistical tests, this algorithm finds the conditional independence relationships among the nodes and uses these relationships as constraints to construct a BN structure (Baesens *et al*., 2002; Frey *et al*., 2003). This algorithm is referred to as a dependency-analysis-based or constraint-based algorithm. The Conditional Independence (CI) test investigates whether two attributes are conditionally independent. There are two common methods to compute the CI test; Pearson chi-square test and log likelihood ratio test (Witten and Frank, 2005). The Likelihood Ratio (LR) tests for target-predictor independence by calculating a ratio between the maximum probability of a result under two different hypotheses. While the Pearson Chi-square

(CHI) asses for target-predictor independence by using a null hypothesis that the relative frequencies of occurrence of observed events follow a specified frequency distribution. MBE explores not only the relations between the class target and predictive attributes, but also the relations among these predictive attributes themselves. Both independence tests; Likelihood ratio and Chi-square have been used to predict and contribute to the proposed ensemble.

Tree Augmented Naive Bayes (TAN): TAN is an improvement over the naive Bayes model as it allows for each attribute to depend on another attribute in addition to the target attribute. The class attribute is the single parent of each node of a NB network: TAN considers adding a second parent to each attribute; the predictive attributes are allowed to point to each other (as long as no cycles are introduced). The decision to add these edges between attributes is made on the basis of a specific goodness of fit measure, such as Maximum Likelihood (ML), Bayesian Dirichlet (BD) (Heckerman *et al*., 1995), Bayesian Information Criterion (BIC) (Grunwald *et al*., 2005), or Akaike Information Criterion (AIC) (Bozdogan, 2000), among others. If the class node and all corresponding edges are excluded from consideration and assuming that there is exactly one node to which a second parent is not added, the resulting classifier has a tree structure rooted at the parentless node. There is an efficient algorithm for finding the set of edges that maximizes the network's likelihood based on computing the network's maximum weighted spanning tree (Witten and Frank, 2005). This method associates a weight to each edge corresponding to the mutual information between the two variables. The TAN learning procedure is as follows:

- Assume the training dataset D, X, Y as input
- Build the tree-like network structure over the predictive attribute X by using the maximum weighting spanning tree
- Add Y as a parent of every X_i where $1 \leq i \leq n$
- Estimate the parameter of TAN (conditional probability of each node given the value of its parents) using ML criterion

When the dataset is small it is preferable to use the BD criterion to prevent the overfitting of the model (Heckerman *et al*., 1995). The proposed ensemble acquires the contribution from TAN classifier with ML test to predict the severity of the breast masses.

Ensemble of Bayesian classifiers: An ensemble of classifiers is a collection of models whose individual predictions are combined by weighted averaging or voting or other majority algorithm. Dietterich (2000) states that "A necessary and sufficient condition for an ensemble of classifiers to be more accurate than any of its individual members is if the classifiers are accurate and diverse". Ensemble algorithms such as bagging, boosting, or random forests enhance the classification performance by combining multiple base classifiers to work as a "team-work" for decision-making (Bauer and Kohavi, 1999). Such team-work approaches not only increase the classification accuracy, but also reduce the chances of overtraining since the team avoids a biased decision by integrating the different predictions from the individual classifiers. The ensemble presented here combines the predictions of three Bayesian classifiers; TAN, MBE with Likelihood independence test and MBE with Pearson Chi-square. Figure 1 shows the component nodes of the proposed ensemble. The ensemble stream is implemented in SPSS Clementine data mining workbench using Intel core 2 Dup CPU with 2.00 GHz (SPSS Clementine 12.0, 2007). SPSS stands for Statistical Package for the Social Sciences.

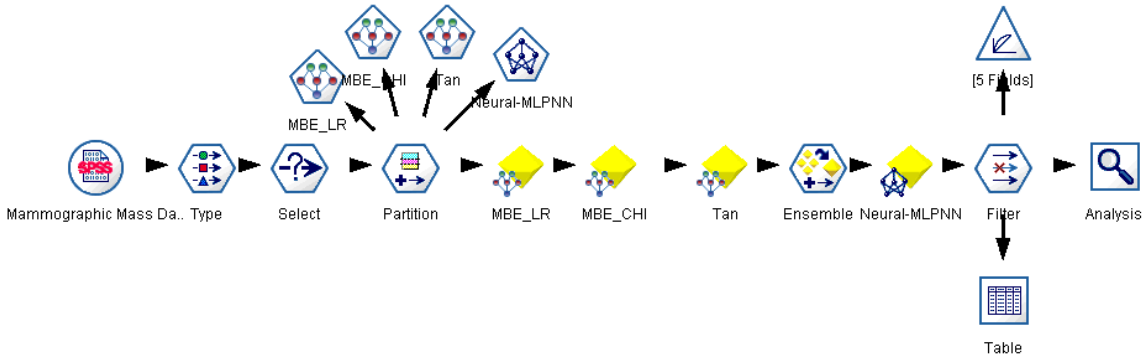

Fig. 1: Data mining stream for the prediction of the severity of breast masses with ensemble of Bayesian networks

Clementine software uses client/server architecture to distribute requests for resource-intensive operations to powerful server software, resulting in faster performance on larger datasets. It is very appropriate as a mining engine with its interface and manipulating modules that allow data examination, manipulation and exploration of any interesting knowledge patterns. The software offers many modeling techniques, such as prediction, classification, segmentation and association detection algorithms. The brief description of each component is given in the following.

Mammographic mass dataset node is connected directly to SPSS file that contains the source data. The dataset was explored for incorrect, inconsistent. Only, the age attribute is normalized and no preprocessing for other attributes. They are ordinal and nominal data types.

Type node specifies the field metadata and properties that are important for modeling and other work in Clementine. These properties include specifying a usage type, setting options for handling missing values, as well as setting the role of an attribute for modeling purposes.

Select node is used to ensure that every sample has a specified class label and discard all samples with undefined ones.

Partition node is used to generate a partition field that splits the data into separate subsets for the training and test the models. In this study, the dataset was partitioned by the ratio 70:30% for training and test subsets respectively.

MBE_LR and MBE_CHI classifier nodes are used to train and test a Bayesian classifier with MBE learning algorithm and Likelihood Ratio (LR) and Pearson Chi-square tests respectively. MBE algorithm selects the set of nodes in the dataset that contain the target attribute's parents, its children and its children's parents. Essentially, MBE identifies all the attributes in the network that are needed to predict the target class. Figure 2a and b illustrate the network topologies with LR and CHI tests respectively. It is clear that there is no direct relation between the class attribute and the mass density in both topologies. It could be concluded that mass density attribute is out of the Markov blanket of the severity class. The MBE model with LR conditional probability test is assumed to be more accurate and experimental results presented here assure this assumption. The predicting accuracy of MBE with Likelihood ratio test is 91.54 and 90.63% of training and test samples respectively. While the same algorithm with Pearson Chi-square test achieves 89.45 and 87.85% predicting success of the same datasets. However, with large datasets there exists may be a processing time-penalty due to the high number of attributes involved. Accuracies are computed as defined later by Eq. 2.

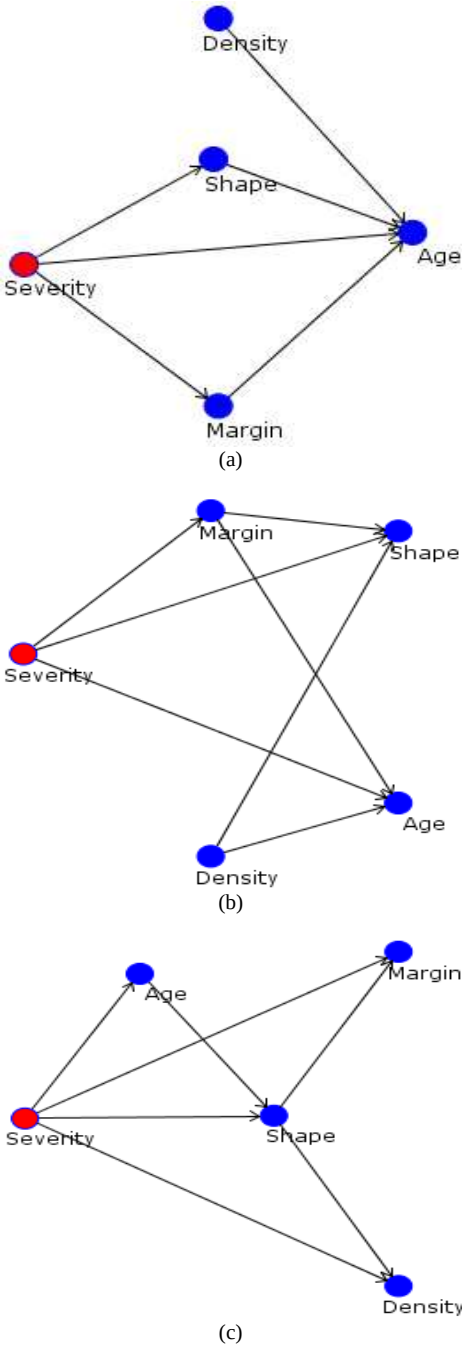

Fig. 2: Topology of the Bayesian networks. (a) MBE learning algorithms with Likelihood ratio test; (b) MBE learning algorithms with Pearson Chi-Square test; (c) TAN learning algorithm

TAN classifier node is to train and test a BN model with TAN learning algorithm where each predictive attributes are allowed to depend on each other in addition to the target attribute, thereby increasing the classification accuracy. In order to prevent overfitting of the classifier, the maximum likelihood is used to control the estimation of the conditional probability for each node given the values of its parents. The TAN classifier achieves 87.07 and 84.72% success of classification the training and test samples respectively.

Ensemble node is used to combine the scored predictions of the three classification models to obtain more accurate results than can be gained from any of the individual models. The proposed system uses confidence scores and the highest confidence wins. However, SPSS Clementine provides variety of majority rules to combine individual predictions including: Voting, confidence-weighted voting and highest confidence wins.

Neural-MLPNN classifier node is trained using the well-known back propagation method with pruning (Thimm *et al*., 1996). It begins with a large network and removes the weakest neurons in the hidden and input layers as training proceeds. The stopping criterion is set based on time; maximum one minute is allowed and the algorithm saved only the network model with the best accuracy achieved. Training MLPNN with pruning method on the mammographic mass dataset for minute has achieved accuracy of 81.13 and 80.90% of training and test samples respectively. Prunning method attains structure of three layers; input, hidden and the output layers with 12, 2 and 1 neuron respectively.

Filter, analysis and evaluation nodes are used to select and rename the classifier outputs in order to compute the performance statistical measures and to graph the evaluation charts.

RESULTS AND DISCUSSION

The performance of each classification model is evaluated using three statistical measures; classification accuracy, sensitivity and specificity. These measures are defined using True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). A true positive decision occurs when the positive prediction of the classifier coincided with a positive prediction of the physician. A true negative decision occurs when both the classifier and the physician suggest the absence of a positive prediction. False positive occurs when the system labels a benign case; a negative one as a positive one (malignant). Finally, false negative occurs when the system labels a positive case as negative (benign). Classification accuracy is defined as the ratio of the number of correctly classified cases and is equal to the sum of TP and TN divided by the total number of cases N:

$$\text{Accuracy} = \frac{TP + TN}{N} \tag{2}$$

Sensitivity refers to the rate of correctly classified positive and is equal to TP divided by the sum of TP and FN. Sensitivity may be referred as a True Positive Rate:

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3}$$

Specificity refers to the rate of correctly classified negative and is equal to the ratio of TN to the sum of TN and FP. False Positive Rate equals (100-specificity):

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{4}$$

The mammographic mass dataset contains 961 sample with class distribution: benign: 516; malignant: 445. There are 162 missing values of different attributes. The whole dataset is divided for training the models and test them by the ratio of 70:30% respectively. The training set is used to estimate each model parameters, while the test set is used to independently assess the individual models. Three models have been trained to predict the severity of breast masses; MBE_LR, MBE_CHI and TAN. These models are applied again to the entire dataset and to any new data. The predictions are combined to build the Bayesian ensemble and compared with the original classes to identify true positives, true negatives, false positives and false negative values. These values have been computed to construct the confusion matrix. The performance is benchmarked with well-known multi-layer neural network.

Table 2 shows the computed confusion matrix, each cell contains the raw number of samples classified for the corresponding combination of desired and actual model outputs. Table 3 presents the values of the statistical parameters (sensitivity, specificity and total classification accuracy) of the predictive models. Sensitivity and Specificity approximates the probability of the positive and negative labels being true. These results show that the sensitivity, specificity and classification accuracy of Bayesian network with MBE learning method and likelihood ratio test are better than those of the other individual classifiers.

Table 2: Confusion matrices of different models of training and test data partitions

Model	Desired output	Training data		Test data	
		Benign	Malignant	Benign	Malignant
Bayesian-MBE_LR	Benign	331	33	133	19
	Malignant	24	285	8	128
Bayesian-MBE_CHI	Benign	327	37	133	19
	Malignant	34	275	16	120
Bayesian-TAN	Benign	305	59	124	28
	Benign	28	281	16	120
Bayesian ensemble (MBE_LR, MBE_CHI, TAN)	Benign	331	33	133	19
	Benign	22	287	8	128
MLPNN	Benign	286	78	117	35
	Malignant	49	260	20	116

Table 3: The values of the statistical measures for different models of training and test data partitions

Model	Partition	Measures (%)		
		Accuracy	Sensitivity	Specificity
Bayesian-MBE_LR	Training	91.53	92.23	90.93
	Test	90.63	94.12	87.50
Bayesian-MBE_CHI	Training	89.45	89.00	89.84
	Test	87.85	88.24	87.50
Bayesian-TAN	Training	87.07	90.94	83.79
	Test	84.72	88.24	81.58
Bayesian ensemble	Training	91.83	92.88	90.93
	Test	90.63	94.12	87.50
MLPNN	Training	81.13	84.14	78.57
	Test	80.90	85.29	76.97

The ensemble classifier has achieved slightly better results for training samples and the same results for test ones. The enhancement in ensemble predictions comes from both MBE_CHI and TAN classifiers; both classifiers give the right prediction with higher confidence.

Gain and Receiver Operating Characteristic (ROC) curves have been used to compare the performances of different predictive models. The gain curves summarize the utility that can be expected by using the respective predictive models, as compared to using baseline information only. Figure 3a shows the cumulative gain curves of the Bayesian models, the proposed ensemble and the neural network for test samples. The higher lines indicate better models, especially on the left side of the chart. The higher curves are of the ensemble and the MBE_LR. ROC procedure is a common way to evaluate the performance of classification models in which the class attribute has two categories by which samples are classified. It is a plot of the sensitivity against one minus the specificity for different values of the threshold. Figure 3b shows the ROC curve of the experimental results. Comparison is usually in terms of the area under the curve, which gives a summary of performance over the whole range of values and is independent of the prevalence of the condition unlike the accuracy, which weights the sensitivity and specificity in proportion to their prevalence.

Table 4: Area under the ROC curve

Model	Area
MBE_LR	0.914
MBE_CHI	0.890
TAN	0.866
Ensemble	0.916
NN	0.813

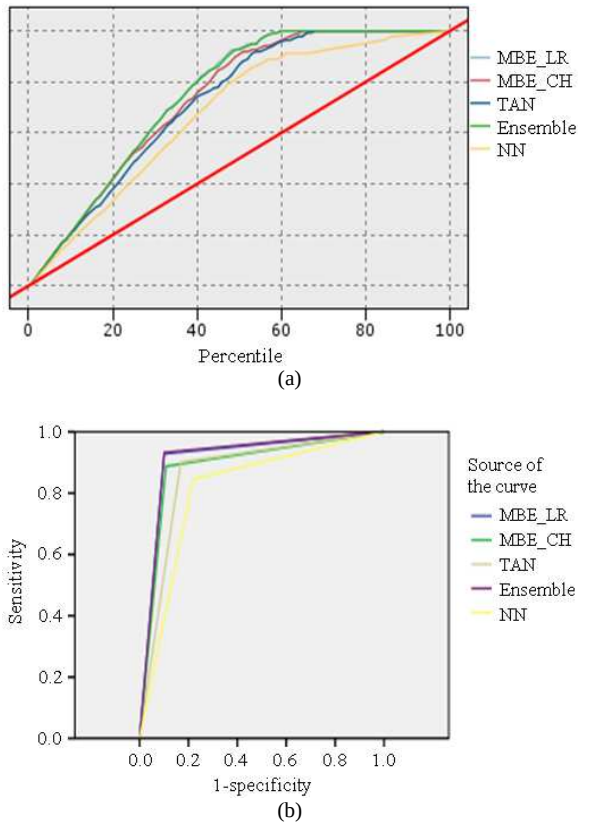

Fig. 3: ROC curve and gain chart for class severity = 1 of all classifiers. (a) gain chart; (b) ROC curve

ROC measures the probability that for any pair of patients, one of whom with an event and one without, the patient for whom the event has occurred will have a higher predicted probability of the event than the other. Table 4 shows the area under the ROC curve for each predicting model. The MBE_LR has the best value among individual models and ensemble has achieved slightly better with 0.916 of ROC area curve.

CONCLUSION

Bayesian network classifiers have three major advantages; they have the ability to deal with missing values, they explicitly provide the conditional

probability distributions of the values of the class attribute given the values of the other input attributes and finally they are easy to comprehend. For these fine proprieties, the awareness to apply and use Bayesian network classifiers in the medical context is increasing. The main goal of this study is to show the effectiveness of these classifiers and their ensemble in the prediction of breast mass severity. Two different implementations of Bayesian network have been applied on the mammographic mass dataset; tree augmented Naive Bayes and Markov blanket estimation learning algorithms. The later may be adapted using the Likelihood ratio test or the Pearson chi-square one. The dataset contains BI-RADS assessment, age, three BI-RADS attributes and type of severity. The dataset has a lot of missing values. The performances of Bayesian classifiers and their ensemble are benchmarked against the multilayer perceptron neural network using statistical measures, gain and ROC charts. Bayesian network classifiers outperformed the multilayer perceptron neural network on the prediction of the severity of breast masses and they provide an elegant way to rank the attributes that most significantly indicate the likelihood of default. On the basis of these results it can be concluded that Bayesian network classifiers may be a competitive alternative to other techniques in medical applications.

REFERENCES

American College of Radiology, 1998. Illustrated Breast Imaging Reporting and Data System (BI-RADS). 3rd Edn., The American College of Radiology, Reston, VA.

Asuncion, A. and D. Newman, 2007. UCI Machine Learning Repository of machine learning databases. University of California, School of Information and Computer Science, Irvine, CA. http://www.ics.uci.edu/~mlearn/MLRepository.html

Baesens, B., M. Egmont-Petersen, R. Castelo and J. Vanthienen, 2002. Learning Bayesian network classifiers for credit scoring using Markov Chain Monte Carlo search. Proceedings of the 16th International Conference on Pattern Recognition (ICPR'02), Aug. 11-15, IEEE Computer Society, Washington DC., USA., pp: 30049. http://portal.acm.org/citation.cfm?id=842819

Bauer, E. and R. Kohavi, 1999. An empirical comparison of voting classification algorithms: Bagging, boosting and variants. Mach. Learn., 36: 105-139. DOI: 10.1023/A:1007515423169

Blake, C. and C. Merz, 1998. UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/MLRepository.html

Bozdogan, H., 2000. Akaike's information criterion and recent developments in information complexity. J. Math. Psychol., 44: 62-91. DOI: 10.1006/jmps.1999.1277

Cheng, J., B. David and W. Liu, 1998. Learning Bayesian Networks from Data: An Efficient Approach based on Information Theory. http://www.cs.ualberta.ca/~jcheng/Doc/report98.pdf

Choua, S., T. Leeb, Y. Shaoc and I. Chenb, 2004. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst. Appli., 27: 133-142. DOI: 10.1016/j.eswa.2003.12.013

Cooper, G., 1999. An Overview of the Representation and Discovery of Causal Relationships Using Bayesian Networks. Computation, Causation and Discovery. AAAI Press/MIT Press, Cambridge, MA., pp: 3-62.

Cruz-Ramirez, N., H. Acosta-Mesaa, H. Carrillo-Calvet and R. Barrientos-Martineza, 2009. Discovering interobserver variability in the cytodiagnosis of breast cancer using decision trees and Bayesian networks. Applied Soft Comput., 9: 1331-1342. DOI: 10.1016/j.asoc.2009.05.004

Cruz-Ramirez, N., H. Acosta-Mesaa, H. Carrillo-Calvet, L. Nava-Fernandezc and R. Barrientos-Martineza, 2007. Diagnosis of breast cancer using Bayesian networks: A case study. Comput. Biol. Med., 37: 1553-1564. DOI: 10.1016/j.compbiomed.2007.02.003

Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensemble of decision trees: Bagging, boosting and randomization. Mach. Learn., 40: 139-158. DOI: 10.1023/A:1007607513941.

Elmore, J., M. Wells, M. Carol, H. Lee, D. Howard and A. Feinstein, 1994. Variability in radiologists' interpretation of mammograms. N. Engl. J. Med., 331: 1493-1499.

Elter, M., R. Schulz-Wendtl and T. Wittenberg, 2007. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Med. Phys., 34: 4164-4172. PMID: 18072480

Frey, L., D. Fisher, I. Tsamardinos, C. Aliferis and A. Statnikov, 2003. Identifying Markov blankets with decision tree induction. Proceeding of 3rd IEEE International Conference on Data Mining, (ICDA'03), IEEE Computer Society Washington, DC, USA., pp: 59-66.

Friedman, N., D. Geiger and M. Goldszmidt, 1997. Bayesian network classifiers. Mach. Learn., 29: 131-163. ftp://ftp.ics.uci.edu/pub/machine-learning-databases

Grunwald, P., I.J. Myung and M.A. Pitt, 2005. Advances in Minimum Description Length: Theory and Applications. MIT Press, USA., ISBN: 10: 0-262-07262-9, pp: 414.

Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, CA., ISBN: 10: 1-55860-901-6.

Heckerman, D., D. Geiger and D.M. Chickering, 1995. Learning Bayesian networks: The combination of knowledge and statistical data. Mach. Learn., 20: 197-243. DOI: 10.1007/BF00994016.

Johnson, R. and D. Wichern, 2002. Applied Multivariate Statistical Analysis. 5th Edn., Prentice-Hall, New York, USA., ISBN: 10: 0130925535, pp: 767.

Kovalerchuck, B., E. Triantaphyllou, J. Ruiz and J. Clayton, 1997. Fuzzy logic in computer-aided breast-cancer diagnosis: Analysis of lobulation. Artif. Intel. Med., 11: 75-85.

Larose, D., 2006. Data Mining Methods and Models. John Wiley and Sons, New York, USA., ISBN: 0471666564, pp: 227-229.

Linda, C., S. Renooij, A. Feelders, A. Groote and M. Eijkemans *et al*., 2008. Technical Report, UU-CS-2008-015, Department of information and computing sciences, Utrecht University, Utrecht, Netherlands. http://www.cs.uu.nl/research/techreps/repo/CS-2008/2008-015.pdf

Nisbet, R., J. Elder and G. Miner, 2009. Handbook of Statistical Analysis and Data Mining Applications. Academic Press, Burlington, MA., ISBN: 978-0-12-374765-5, pp: 253-256.

Pendharkar, P., J. Rodger, G. Yaverbaum, N. Herman and M. Benner, 1999. Association statistical, mathematical and neural approaches Aligning Bayesian network classifiers with medical contexts for mining breast cancer patterns. Expert Syst. Applied, 17: 223-232.

Singh, S. and R. Al-Mansoori, 2000. Identification of regions of interest in digital mammograms. J. Intel. Syst., 10: 183-210.

SPSS Clementine 12.0, 2007. Data mining workbench software. Product of SPSS inc. http://www.cad100.net/247_data-mining-workbench-SPSS-Clementine-12.html

Thimm, G., E. Furuhashi and Takeshi, 1996. Neural network pruning and pruning parameters. Proceeding of the 1st Workshop on Soft Computing, Aug. 1996, Idiap Publications Japan, pp: 1-1. http://publications.idiap.ch/index.php/publications/show/1011

Witten, I. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. 2nd Edn., Morgan Kaufmann, San Francisco, CA., ISBN: 10: 0120884070, pp: 260.