AENSI Journals

Australian Journal of Basic and Applied Sciences

ISSN:1991-8178

Journal home page: www.ajbasweb.com

Experimental Analysis Towards Realizing Breast Cancer Prognosis Using Diverse Machine Learning Classifiers

[1]Sandeep Chaurasia, [1]Prasun Chakrabarti, [2]Yu Cheng

[1]Department of Computer Science, School of Engineering, Sir Padampat Singhania University, India
[2]Researcher, SAP Labs, Shanghai, China

ARTICLE INFO

Article history:
Received 2 March 2014
Received in revised form
13 May 2014
Accepted 28 May 2014
Available online 13 June 2014

Keywords:
Breast cancer, classifier, decision tree, Naïve Bayes, Neural network, support vector machine

ABSTRACT

The adequate diagnosis of breast cancer is one of the major challenges in the medical field. Supervised machine learning has been used to simulate a model of the distribution of class label in terms of predictor features. The resultant classifier is then used for helping doctors' forms a secondary opinion for better diagnosis. The performance of various machine learning techniques has been analyzed over the four distinguishes breast cancer data sets. A comparison on the performance of the results has produced among the classifiers as decision tree, Naïve Bayes, Naïve Bayes using kernel, neural network, auto association multi layer Perceptron and support vector machine. The obtained results shows that SVM could classify more accurate when there is no missing data, but with missing data Naïve Bayes using kernel method works fast and generate hypothesis more accurately.

To Cite This Article: Sandeep Chaurasia, Prasun Chakrabarti and Yu Cheng., Experimental Analysis Towards Realizing Breast Cancer Prognosis Using Diverse Machine Learning Classifiers. ***Aust. J. Basic & Appl. Sci.,*** *8(9): 31-37, 2014*

1 Machine Learning in Medical Context – An Introduction:

In recent years the mortality rate of breast cancer has significantly increased. Since 2000 in USA, breast cancer is the second largest cause of cancer deaths among women following the lung cancer. In USA 40,600 deaths from breast cancer in 2009, 400 were men (Chaurasia S., 2012). Mammography is the most successful method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes (Siegel, R., 2013). An effective way to reduce the high mortality rate of breast cancer is to detect it at an early stage. Prevention is still a mystery and the only way to reduce the mortality rate of patients by early detection. Classification techniques are reducing the possible errors that might be made because of unverified experts; provide more detailed medical data for examination in a shorter time (Chaurasia S., and Chakrabarti P., 2013). A study showed that if the cancerous cells are detected before spreading to other organs then the survival rate for patients is more than 97% (American Cancer Society Homepage, http://www.cancer.org/ 2008).

One of the primary goals of machine learning is to device an efficient algorithm for training computers to automatically acquire effective and accurate model from experience. It is providing a technique, method tools that can assist in solving prognosis and diagnosis problems in a variety of medical domains. It is argued that the successful implementation of ML methods can help the integration of computer-based systems in the healthcare environment providing opportunities to facilitate and enhance the work of medical experts and ultimately to improve the efficiency and quality of medical care. The next section summarizes some major ML application areas in medicine.

2 Related Work:

There has been research with WBCD the breast cancer database on computer aided diagnosis and prognosis of breast cancer. Quinlan J.R has presented an algorithm using C4.5 decision tree method using 10-fold cross validation and reported an accu-racy of 94.74% (Quinlan, J.R., 1996). Hamilton, Shan and Cercone presented a method named Rule induction algorithm based on approximate classifications and reported the accuracy of 94.99% (Hamiton, H.J., 1996). Nauck D, and Kruse R presented an neuron-fuzzy technique for classification of medical data and reported the accuracy of 95.06% (Nauck, D., R. Kruse, 1999). Abonyi and Szeifet used the application of the supervised fuzzy clustering technique and reported an accuracy of 95.57% (Abonyi, J., F. Szeifert, 2003). Albercht, Lappas, Vinterbo, Wong and Ohno-Machado presented a learning algorithm that combined logarithmic simulated annealing with the perceptron algorithm was used and reported an accuracy of

Corresponding Author: Sandeep Chaurasia, Department of Computer Science, School of Engineering, Sir Padampat Singhania University, India.
E-mail: chaurasia.sandeep@gmail.com

95.57% (Albrecht, A.A., 2002). Guijarro-Berdias B, Fontenla-Romero O, Perez-Sanchez B, and Fraguela P presented a learning algorithm by applying linear- least squares method and reported an accuracy of 96% (Guijarro-Berdias, B., 2007). Karabatak and Cevdet-Ince presented an automatic diagnosis system for detecting breast cancer based on Association Rules (AR) and neural networks (NNs), and reported an accuracy of 97.4% (Karabatak, M., M. Cevdet-Ince, 2009).

3 Brief Reviews of Classification Techniques:

The classifier's evaluation is most often based on prediction accuracy. There are various methods to evaluate the accuracy by dividing the data set, two-third for training and one-third for testing. Another method as cross validation, the training set is split into mutually exclusive and equal sized subset and for each subset the classifier is trained on the union of all others subset. And the last is leave-one-out validation is a modified case of cross validation (Kotsiantis, S.B., 2007). A large number of techniques have been developed based on logic/symbolic based techniques, perceptron based techniques and statically bayesian network.

3.1 Logic and Symbolic based:

Decision trees: Decision tree are trees that classify instances by sorting them based on feature value in an instance to be classified, each branch represents a value that the node can assume. Instances are classified starting at the root node and sorted based on their feature value (Murthy, 1998). C4.5 is a well known algorithm and is an extension of Quinlan's earlier ID3 algorithm having a very good combination of error rate and speed (Quinlan, J.R., 1993).

3.2 Statistical Learning:

Naïve Bayes classifier: Naïve Bayes are simple Bayesian network composed of DAG (direct acyclic graph) with one parent and several children. But the assumption of independence among child nodes is clearly almost always wrong and for this reason the quick learner Naïve Bayes classifier are usually less accurate than other sophisticated learning algorithm. Conditional probabilities: $Pi(xi|C = c)$, the probability that the feature value in the *i*-th position is equal to *xi* given class *c*, were estimated using KDE from a set of labeled training data (*X*, *C*). KDE is a non-parametric way of estimating the probability density function population (Parzen, E., 1962). The probability $P_i(x_i|C = c)$was estimated using Equation.

$$P_i(x_i|C = c) = \frac{1}{N_c h}\sum_{j=1}^{N_c} K\left(x_i, x_{j\,|i|c}\right) \qquad K(a,b) = \frac{1}{\sqrt{2\pi}} e^{\frac{(a-b)^2}{2h^2}}$$

where K is a Gaussian function kernel with mean zero and variance 1, N_c is the number of the input data *X* belonging to class *c*, *xj*|*i*|*c* is the feature value in the *i*-th position of the *j*-th input $X = (x_1\ x_2\ \ldots\ x_i\ \ldots\ x_n)$ in class c, and *h* is a bandwidth, or a smoothing parameter. To optimally estimate the conditional probabilities, *h* was optimized on the training data set.

3.3 Perceptron based Techniques:

Artificial neural network: (ANN) depends on input data, its activation function and weight of each input connection. There are several algorithms by which a network can be trained (Neocleous, C., C. Schizas, 2002), but the most popular algorithm is back propogation (BP) algorithm. The back propogation algorithm will perform a number of weight modifications before it concludes with a good weight configuration for n training instances and w weights each epoch in learning takes O (nw) time. Bourlard and Kamp (1988) present paper for auto-association, the nonlinearities of the hidden units are useless and that the optimal parameter values can be derived directly by purely linear techniques relying on singular value decomposition and low rank matrix approximation, similar in spirit to the well-known Karhunen-Loève transform.

3.4 Support Vector Machine:

The support vector machine is originally a binary classification method developed by Vapnik et.al at Bell laboratories (Vapnik, V., 1999). For a binary problem, we have training data point $\{X_i Y_i\}, i = 1...l,\ y_i \in \{-1,1\}, x_i \in R^d$. Suppose we have some hyperplane that separates or classify the positive label from the negative label with a separating hyperplane. The points x which is on the hyperplane satisfy w·x + b = 0, where w is normal to the hyperplane |b|/||w||, is the perpendicular distance from the hyperplane to the origin, and ||w|| is the Euclidean norm of w. For non linear separable case, selection of appropriate kernel function is important because kernel function is responsible to transformed feature space in which training set instance will be classified. Training the SVM is performed by solving nth dimensional QP problem, where N is the number of samples in training dataset which involves large matrix operations.

In this paper, the four breast cancer dataset has been analyzed over different machine learning principles of classification techniques. This paper is organized as follows, section 2 provides the brief of the related work

done, section this section highlighted a brief introduction of various classification techniques and algorithm, section 4 provide a detailed description of data sets, section 5 shows the comparison statistics of the mention techniques with the acquired results. Finally section 6 concludes the result.

4 Breast Cancer Dataset Overview:

4.1 Breast cancer data set 1:

The WBCD dataset consists of 699 instances taken from Fine Needle Aspirates (FNA) of human breast tissue. Each record in the database has nine attribute.

Table 1: Attribute of data set 1.

Attribute	Value of the attribute
1. Sample code number	id number
2. Clump Thickness	1 – 10
3. Uniformity of Cell Size	1 – 10
4. Uniformity of Cell Shape	1 – 10
5. Marginal Adhesion	1 – 10
6. Single Epithelial Cell Size	1 – 10
7. Bare Nuclei	1 – 10
8. Bland Chromatin	1 – 10
9. Normal Nucleoli	1 – 10
10. Mitoses	1 – 10
11. Class:	2 - benign,4- malignant

N = 683 observation: class distribution: benign: 444; malignant: 239

4.2 Breast cancer data set 2:

This data set can be used to predict the severity (benign or malignant) of a mammo-graphic mass lesion from BI-RADS attributes and the patient's age. It contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 516 benign and 445 malignant masses.

Table 2: Attribute of data set 2.

Attribute	Attribute description	Value of attribute
1	BI-RADS assessment	1 to 5
2	Age: patient's age in years	Integer
3	Shape: mass shape	round=1 oval=2 lobular=3 irregular=4
4	Margin: mass margin	circumscribed=1 microlobulated=2 ob-scured=3 ill-defined=4 spiculated=5
5	Density : mass	high=1 iso=2 low=3 fat-containing=4
6	Severity:	benign=0 or malignant=1

N = 961 observation: class distribution: benign: 516; malignant: 445

4.3 Breast cancer data set 3:

The first 30 features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. Total number of instances are198 and total number of attributes are 34 (ID, outcome, 32 real-valued input features). If recurrences before 24 months then it is positive(R) or if non-recurrence (N) beyond 24 months then it is negative.

Table 3: Attribute of data set 3 1) ID number.

2) Outcome (R = recurring , N = nonrecurring)
3) Time (recurrence time if field 2 = R, disease-free time if field 2 = N)
4-33) Ten real-valued features are computed for each cell nucleus:
- a) radius (mean of distances from center to points on the perimeter)
- b) texture (standard deviation of gray-scale values)
- c) perimeter
- d) area
- e) smoothness (local variation in radius lengths)
- f) compactness (perimeter^2 / area - 1.0)
- g) concavity (severity of concave portions of the contour)
- h) concave points (number of concave portions of the contour)
- i) symmetry
- j) fractal dimension ("coastline approximation" - 1)

4.4 Breast cancer data set 4:

The total number of instances are 569 and total number of attributes are 34 (ID, diag-nosis and 32 real-valued input features). The diagnosis is classified into two labels malignant and benign with distribution on 212 and 357 respectively.

Table 4: Attribute of data set 4.

1) ID number 2) Diagnosis (M = malignant, B = benign) 3-32) Ten real-valued features are computed for each cell nucleus: a) radius (mean of distances from center to points on the perimeter) b) texture (standard deviation of gray-scale values) c) perimeter d) area e) smoothness (local variation in radius lengths) f) compactness (perimeter^2 / area - 1.0) g) concavity (severity of concave portions of the contour) h) Concave points (number of concave portions of the contour) i) symmetry j) fractal dimension ("coastline approximation" - 1) N = 569 observation; class distribution: Benign: 357; Malignant: 212

5 Results and Discussions

The original data is present in the form of analogue values with different range of data. The data are converted to their equivalent integer or real number form. Then the mean and the standard deviation are calculated to normalize the data. Then the label field is identified for dataset 1 it is value 2 for benign and 4 for malignant, in dataset 2 it is 0 for benign and 1 for malignant. For dataset 3 it is B for benign and M for malignant and dataset 4 it is N for benign and R for malignant. If any missing data is encountered then the missing value is replaced by the average value of the column. To measure the performance of the breast cancer diagnosis of the classifiers used in this investigation we have divide the evaluation it into two parts first is to determine per-formance result accuracies by means of classification accuracy (Marcano-Cedeño, A., 2011), analysis of speci-ficity and sensitivity, and confusion matrix and the second is by the performance results in term of ROC, related to ROC curve analysis and area under the curve (AUC). We explain the methods used in each part in the following sections.

5.1 Performance Evaluation:

***Classification accuracy*.** In this study the classification accuracy for each data sets are calculated using the following equation:

$$\text{accuracy}=\text{TP+TN}\Big/(TP+TN+FP+FN)$$

where TP, TN, FP and FN denotes true positive, true negative, false positive and false negative.

Sensitivity and specificity:

For measuring performance by means of sensitivity and specificity analysis, we use the following expressions.

$$\text{Sensitivity}=\frac{\text{TP}}{\text{TP+FN}}(\%):\text{Sensitivity}=\frac{\text{TN}}{\text{FP+TN}}(\%)$$

Performance results ROC:

Receiver Operating Characteristic (ROC) curve is a graphical plot true positive rate vs. the fraction of false positives out of the total actual negatives i.e. false positive rate. ROC is widely used in biomedical research to assess the performance of diagnos-tic tests (Witten, I.H., E. Frank, 1999). The Area under the ROC Curve (AUC): Another method of evaluating classifier performance is the area under ROC curve (AUC). The AUC close to 1 indicates very reliable diagnostic test (Bradley, A.P., 1997).

Table 5: Confusion matrices of classifiers used for classification of breast cancer data set 1 & 2.

Type classifier	Desired result	Output results data set 1		Output results data set 2	
		Positive	Negative	Positive	Negative
Decision tree	Benign record	429	15	287	45
	Malignant record	15	224	158	470
Naïve bayes	Benign record	424	6	332	83
	Malignant record	20	233	73	345
Naïve bayes kernel	Benign record	424	2	372	96
	Malignant record	20	237	72	420
Neural Net	Benign record	430	9	353	90
	Malignant record	14	230	92	426
Auto MLP	Benign record	428	6	359	94
	Malignant record	16	233	86	422
SVM	Benign record	237	32	330	66
	Malignant record	2	412	73	361

Table 6:. Classification accuracies of classifiers used for detection of breast cancer data set1.

Type classifier	Classification accuracies (%)		
	Specificity	Sensitivity	Total classification accuracy
Decision tree	96.62	93.72	95.61
Naïve bayes	98.60	92.09	96.19
Naïve bayes kernel	99.5	92.21	96.78
Neural net	97.9	94.2	96.64
Auto MLP	98.6	93.5	96.78
SVM	92.79	99.16	95.02

Table 7: Classification accuracies of classifiers used for detection of breast cancer data set 2.

Type classifier	Classification accuracies (%)		
	Specificity	Sensitivity	Total classification accuracy
Decision tree	74.84	86.44	78.79
Naïve bayes	82.53	80	81.27
Naïve bayes kernel	85.36	79.48	82.41
Neural net	82.23	79.68	81.56
Auto MLP	83.07	79.24	81.27
SVM	84.54	81.88	83.25

Table 8: Confusion matrices of classifiers used for classification of breast cancer data set 3 & 4.

Type classifier	Desired result	Output results data set 3		Output result data set 4	
		Positive	Negative	Positive	Negative
Decision tree	Benign record	151	47	193	12
	Malignant record	0	0	19	345
Naïve bayes	Benign record	108	27	190	15
	Malignant record	20	43	22	342
Naïve bayes kernel	Benign record	137	37	190	16
	Malignant record	10	14	20	341
Neural Net	Benign record	129	26	199	8
	Malignant record	21	22	13	349
Auto MLP	Benign record	131	29	198	5
	Malignant record	18	20	14	352
SVM	Benign record	144	34	204	22
	Malignant record	13	7	8	335

Table 9: Classification accuracies of classifiers used for detection of breast cancer data set 3.

Type classifier	Classification accuracies (%)		
	Specificity	Sensitivity	Total classification accuracy
Decision tree	NaN	76.26	76.29
Naïve bayes	68.2	80	64.58
Naïve bayes kernel	58	78.73	74.18
Neural net	51.16	83.22	75.71
Auto MLP	52.63	81.8	75.32
SVM	65	80.8	75

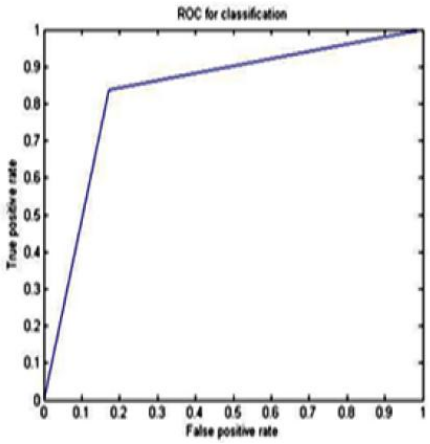

Fig. 1: ROC for support vector machine for data set 2 with AUC = 0.8419.

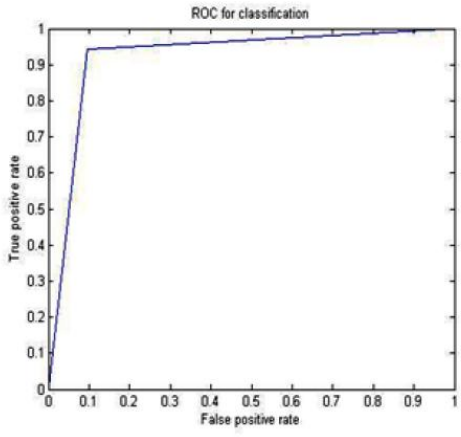

Fig. 2: ROC for support vector machine for data set 4 with AUC = 0.9246

Table 10: Classification accuracies of classifiers used for detection of breast cancer data set 4.

Type classifier	Classification accuracies (%)		
	Specificity	Sensitivity	Total classification accuracy
Decision tree	94.78	94.14	94.56
Naïve bayes	93.95	92.68	93.51
Naïve bayes kernel	94.45	92.23	93.66
Neural net	96.4	96.13	96.31
Auto MLP	96.17	97.5	96.66
SVM	93.83	96.22	95.75

6 Conclusion:

In this study among the four data sets, the accuracy of Naïve Bayes using kernel method, neural network and support vector machine are comparable, though the accuracy depends on some factors such as missing data, number of data inputs and distribution of labels across the feature vectors. In data set 1, the ability of Naïve Bayes kernel method to handle missing data as a quick learner classify more accurately than support vector machine and neural network with an accuracy of 96.78% and with roc 0.997. In dataset 2 the number of observation increased by 33% as a value of 961 without any missing data, the support vector machine is best classifier with an accuracy of 83.25% and roc is 0.8419. In data set 3 the class 2 is not identified because of less number of observation and classes are not distributed evenly so the decision tree is identifying only a single class. For data set 3 neural networks works more accurately with an accuracy of 75.71 and roc is 0.709. In data set 4 without any missing data the accuracy of neural network and support vector machine is comparable with 96.66 and 95.75.

REFERENCES

Abonyi, J., F. Szeifert, 2003. Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recognition Letters.

Albrecht, A.A., G. Lappas, S.A. Vinterbo, C.K. Wong, L. Ohno-Machado, 2002. Two applications of the LSA machine. Proceedings of the 9th international conference on neural information processing, 184-189.

American Cancer Society Hompage, http://www.cancer.org/ 2008

Bourlard, H., Y. Kamp, 1988. Auto-association by multilayer perceptrons and singular value de-composition Biological Cybernetics, 59(4-5): 291-294.

Bradley, A.P., 1997. The use of the area under the ROC curve in the evaluation of ma-chine learning algorithms. Pattern Recognition, 30(7): 1145-1159.

Chaurasia S. and Chakrabarti P., 2013. "An Approach with Support Vector Machine using Variable Features Selection on Breast Cancer Prognosis" International Journal of Advanced Research in Artificial Intelligence(IJARAI), 2(9).

Chaurasia S., Chakrabarti P. and Chourasia N., 2012. Article: An Application of Classification Techniques on Breast Cancer Prognosis. International Journal of Computer Applications. 59(3): 6-10.

Guijarro-Berdias, B., O. Fontenla-Romero, B. Perez-Sanchez, P. Fraguela, 2007. A linear learning method for multilayer perceptrons using leastsquares. Lecture Notes in Computer Science, 365-374. 10.1007/978-3-540-77226-238.

Hamiton, H.J., N. Shan, N. Cercone, 1996. RIAC: A rule induction algorithm based on approximate classification. In International conference on engineering applications of neural networks, University of Regina.

Karabatak, M., M. Cevdet-Ince, 2009. An expert system for detection of breast cancer based on association rules and neural network. Expert Systems with Applications, 36: 3465-3469.

Kotsiantis, S.B., 2007. Supervised machine learning: A review of classification techniques, 2007. Informatica, 31: 249-268

Marcano-Cedeño, A., J. Quintanilla-Domínguez, D. Andina, 2011. WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Ap-plications, 38(8): 9573-9579, ISSN 0957-4174.

Murthy, 1998. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey, Data Mining and Knowledge Discovery, 2: 345-389.

Nasseer M. Basheer, Mustafa H. Mohammed, Classification of Breast Masses in Digital Mammograms Using Support Vector Machines. International Journal of Advanced Re-search in Computer Science and Software Engineering, 3(10).

Nauck, D., R. Kruse, 1999. Obtaining interpretable fuzzy classification rules from medical data. Artificial Intelligence in Medicine, 16: 149-169.

Neocleous, C., C. Schizas, 2002. Artificial Neural Network Learning: A Comparative Review, LNAI 2308, pp: 300-313, Springer-Verlag Berlin Heidelberg.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat., 33: 1065-1076.

Quinlan, J.R., 1993. C4.5: Programs for machine learning. Morgan Kaufmann, San Fran-cisco.

Quinlan, J.R., 1996. Improved use of continuous attributes in C4.5. Journal of Artificial Intelligence Research, 4: 77909.

Siegel, R., D. Naishadham, A. Jemal, 2013. "Cancer Statistics, 2013", CA: A Cancer Journal for Clinicians, 63(1): 1-30.

Vapnik, V., 1999. The nature of statistical learning Theory, 2nd Ed. Springer, NewYork.

Witten, I.H., E. Frank, 1999. Data mining: Practical machine learning tools and tech-niques with java implementations. San Francisco: Morgan Kaufmann Publishers, pp: 89-97, pp: 125-127, pp: 159-161.