Analysis of Classification Algorithms towards Breast Tissue Data Set

I. Ravi

Assistant Professor, Department of Computer Science,

K.R. College of Arts and Science, Kovilpatti, Tamilnadu, India

Abstract

Data mining techniques provides a new direction to extract information from the massive databases. Popular data mining techniques are clustering, classification, association analysis, regression, summarization, time series analysis and sequence analysis, etc. Classification technique is one of the most conventional and popular data mining techniques which are used to classify the data and extract the information. In this research work, breast tissue dataset is used for performing classification task. This classification task helps to classify the data set into six classes which gives the breast tissue type. In this work we converse about three different classifiers such as Naive Bayes, IBK (Instance Based K- Nearest Neighbor) and J48 to process the Breast Tissue dataset and recognize the significance classification of test data using WEKA (Waikato Environment for Knowledge Analysis) tool. Classification accuracy, error rate and execution time are used in this comparative analysis. From the results, it is observed that the J48 algorithm efficiency is better than other algorithms.

I. Introduction

Data mining is an essential step of knowledge discovery process by analyzing the massive volumes of data from various perspectives and summarizing it into useful information [1]. Data mining is used in numerous applications such as medical, stock analysis, fault analysis, forecasting, and science examination. There are numerous task accomplished by data mining they are classification, clustering, association rule mining, prediction, outlier analysis, time series. Classification is one of the most overlooked and efficient task. Data mining in cancer research has been one of the important research topics in biomedical science during the recent years [4].

In medical field data mining tasks work more swiftly than the before years. Breast cancer is one of the most invasive parts found between women and provides the path to increase the output and cut down the cost. More than one million cases are affected and nearly 600,000 deaths occurring worldwide annually [2]. Cancer is a disease and it is characterized as uncontrolled growth and spread of the abnormal cells and the capability to invade other tissues that can be caused by both external factors like radiation and internal factors like hormones [3]. Cancer is one which invades from the cells which are divisible and grow uncontrollably. According to the survey of United States in 2014 there are 232,670 females and 2,360 males having this type of breast cancer [10]. Among them 40,000 females and 430 males were died during the period this survey [10].

The classification algorithm is used to provide the accuracy of classification using the correctly classified instances. Performance measure is calculated by TP, FP rate and error rate. The three classifier algorithms, naïve bayes, IBK, J48 are compared to find the best algorithm among these. Comparative analysis is done by WEKA

(Waikato Environment for Knowledge Analysis) tool and the dataset breast tissue is collected from the UCI repository. A major class of problems in medical science involves the diagnosis of disease based upon various tests performed upon the patient. The classifier system in medical diagnosis is increasing gradually [9].

The classification of breast cancer data can be useful to predict the outcome of some diseases or discover the genetic behavior of tumors and there are many techniques to predict and classification breast cancer pattern [5][8].

II. Literature review

Dursen Delen et.al.,[5] have predicted the breast cancer survivability and analyzed the comparison between neural networks, decision tree induction, logistic regression classifiers to calculate accuracy, sensitivity, specificity, confusion matrix and to predict the person who survive. Authors have collected the breast cancer dataset from Seer database and used WEKA tool to find the accuracy.

K.R Lakshmi et.al., **[**6**]** have analyzed the comparative study between the classifiers such as SVM, PNN, k-NN, BLR, MLR, PLS-DA, PLS-LDA to calculate the accuracy, error rate and precisions, performance and find out which one is best algorithm among these. Here author have used the Tanagra tool and the dataset of breast cancer from seer.

A.Priyanga et.al.,[7] have described the comparative study on cancer prediction system based on the three classifiers such as decision tree J48, ID3, Naïve bayes to calculate accuracy, the range of risks have been determined by four values such as very low, low, high, very high and the prediction is validated. Author used the breast cancer dataset from seer and found the accuracy values using WEKA tool.

Vikas Chaurasia et.al.,[8] have analyzed the comparative study between the classifiers such as SMO, IBK, BF Tree to find the performance, accuracy, simulation result, comparison between parameter, average rank and efficient algorithm is found. The author have collected the breast cancer data from the UC Irvine machine learning repository and found the simulating result in WEKA 3.6.9.

G.RaviKumar et.al.,[9] have analyzed the study of comparison between J48, Naive bayes, KNN, SVM, MLP, Logistic to find the performance and accuracy, error rate and execution time and the effective algorithm is found. Author have collected the Wisconsin breast cancer data from the UCI repository dataset and produced the result in WEKA.

S.Aruna et.al [11] have compared the performance of supervised learning algorithm and used naïve bayes, SVM, Radical basis neural network, Decision tree, J48, Simple cart and found the quality of the classifier for detecting the disease. Implemented in WEKA and the datasets WBC, WDBC, Puma diabetes and Breast tissue are used and these are collected from UCI repository. SVM have produced the best accuracy result compared with another.

Endo et al., [3] implemented machine learning algorithms to predict survival rate of breast cancer patient. Author has collected the data from the SEER dataset with high rate of positive. And author found that logistic regression had the highest accuracy and J48 decision trees model has best sensitivity.

Dr. S.Vijayarani et al., [13] analyses the performance of different classification function techniques in data mining for predicting the heart disease from the heart disease dataset. The performance factors used for analysis are accuracy and error rate.

III. Methodology

To extract data from large set of database, information retrieval can be used. By using bayes, lazy and trees classification we find the best algorithm for effective information retrieval of breast tissue data set. The process flow of comparative analysis is illustrated in fig.1.

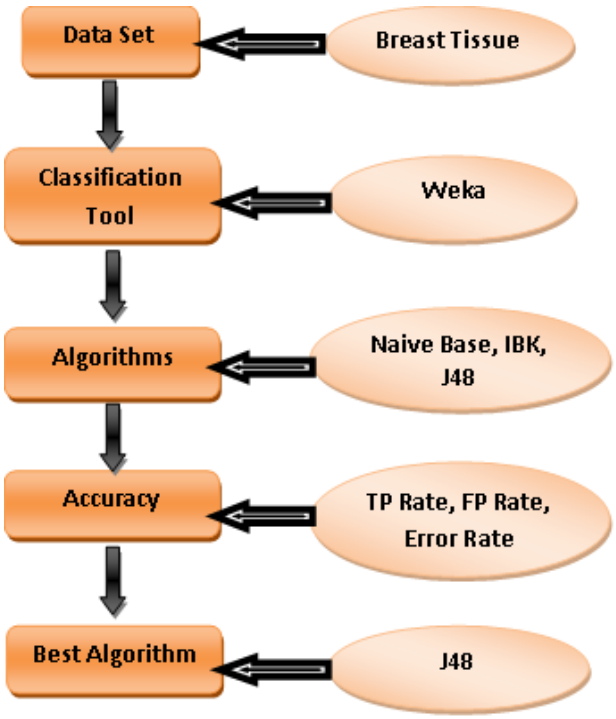

Fig 1: Flow diagram of comparative analysis

A. Dataset

The Breast Tissue dataset is collected from the UCI (UC Irvine) repository. This dataset contains 106 instances and 10 attributes. The machine learning data mining tool called WEKA (Waikato Environment for Knowledge Analysis) is used to assess the performance of three different classification algorithms [14].

B. An Overview of Breast Tissue:

Breasts have two types of tissues: glandular tissues and supporting (stromal) tissues [15] which are mainly function to make milk for breastfeeding. The glandular part of the breast includes the lobules and ducts. Women who are breastfeeding, the cells of the lobules make milk. The milk then moves through the ducts-tiny tubes that carry milk to the nipple [15] and each breast has several ducts that lead out to the nipple. The supporting tissues of the breast have fatty tissue and fibrous connective tissue that give the breast its size and shape. Any of these parts of the breast can undergo changes that cause symptoms. These breast changes can be either benign breast conditions or breast cancers[15].

C. Classification:

Classification is a data mining practice used to predict group membership for data instances. In order to calculate the result, the algorithm implements a training set including a set of attributes and the respective result, usually called prediction attribute. In this research we have analysed three classifiers namely bayes, lazy and

trees to predict which of the algorithm is most suitable for Breast Tissue dataset. In bayes we measure the naive bayes, in lazy we measure the Instance Based K- Nearest Neighbor (IBK) and in trees we measure the J48.

D. Naive Bayes

The Bayesian classification acts as a probabilistic [16] learning method. Naive bayes classifiers are the most successful algorithm for learning to classify text documents. Naive bayes is based on the Bayesian theorm. It is mostly suited when the dimensionality of the inputs is high. The advantage of naive bayes is it requires a small amount of training data to estimate the parameters.

E. IBK (Instance based K-Nearest Neighbor):

The IBK algorithm is a K-Nearest–Neighbor classifier is a non-parametric [17] method used for classification and regression. K-NN is a type of instance-based learning, or lazy learning. The set of objects or the object property value (for k-NN regression) can be taken as neighbors. This can be thought of as the training set for the algorithm, though no open and clear training step is required.

Input:
D//**Training data**
K// **Number of neighbors**
T// **Input tuple to classify**
Output:
C
// **Class to which t is assigned**
IBK Algorithm:
//**Algorithm to classify tuple using IBK**
$N=\emptyset$
// **Find set of neighbors, N, for t**
for each $d \in D$ **do**
If $|N| \leq K$ *then*
$N=N \cup d$;
else
If $\exists u \in N$ **such that** $sim(t,u) \geq sim(t,d)$ **then**
begin
$N=N-u$;
$N=N \cup d$;
end
//**Find class for classification**
C=**class to which the most** $u \in N$ **are classified;**

F. J48

J48 is an open source java implementation of the C4.5 algorithm [18] in the weka data mining tool. C4.5 is an algorithm developed by Ross Quinlan which is used to generate a decision tree for the purpose of classification, and for this reason, C4.5 is frequently referred to as a statistical classifier.

J48 generate decision trees [19] from a set of categorized training data using the concept of information entropy. By splitting the data into smaller subsets, each attribute of the data can be used to make a decision. J48 examines the information gain that results from choosing an attribute for splitting the data. To build the conclusion, the attribute with the highest normalized information gain is used and then the algorithm recurs on the smaller subsets [19]. If all instances in a subset belong to the same class, the splitting procedure stops. A leaf node is shaped in the decision tree telling to choose that class.

IV. Experimental Results

In this paper we used 10-fold cross-validation method to estimate the performance of three different classification methods. We used Breast Tissue dataset which has 106 instances and 10 attributes.

A. Accuracy Measure and Error Rate

The term accuracy refers to the correctly classified instances by the total number of instances present in the dataset.

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN}$$

Where TP- True Positive, FP- False Positive, TN- True Negative, FN- False Negative.

TP Rate is the ability which is used to find the high true-positive rate . The true-positive rate is also called as sensitivity.

$$TPR = \frac{TP}{TP+FN}$$

$$FPR = \frac{FP}{TN+FP}$$

Precision is the ratio of modules correctly classified to the number of entire modules classified fault-prone. It is proportion of units correctly predicted as faulty.

$$Precision = \frac{TP}{TP+FP}$$

F- Measure is the one has the combination of both precision and recall which is used to compute the score. This kind of measure is frequently used in the field of Information Retrieval to calculate the query classification performance.

$$F - Measure = \frac{2*Recall*Precision}{Recall+Precision}$$

The mean absolute error (MAE) is defined as the quantity used to measure how close predictions are to the eventual outcomes. It measures the accuracy for the random and continues variables. The root mean square error (RMSE) is defined as commonly used compute of the differences between values predicted by a model and the values actually observed. It is a good measure of accuracy. Relative error is a measure of the uncertainty of measurement compared to the size of the measurement [13]. The root relative squared error is computed by dividing the root mean square error (RMSE).The accuracy measure for three classification algorithms is shown in Table 1.

Table 1: Accuracy Measure for given three Classifiers.

The following fig.2 illustrates the accuracy measure using the algorithms Naive base, IBK and J48. From the chart we declare that the J48 act best according to correctly classified instances.

Algorithm	Correctly classified instances (%)	Incorrectly classified instances (%)
Naive Bayes	94.34 %	5.67 %
IBK	91.51 %	8.50 %
J48	95.28 %	4.71 %

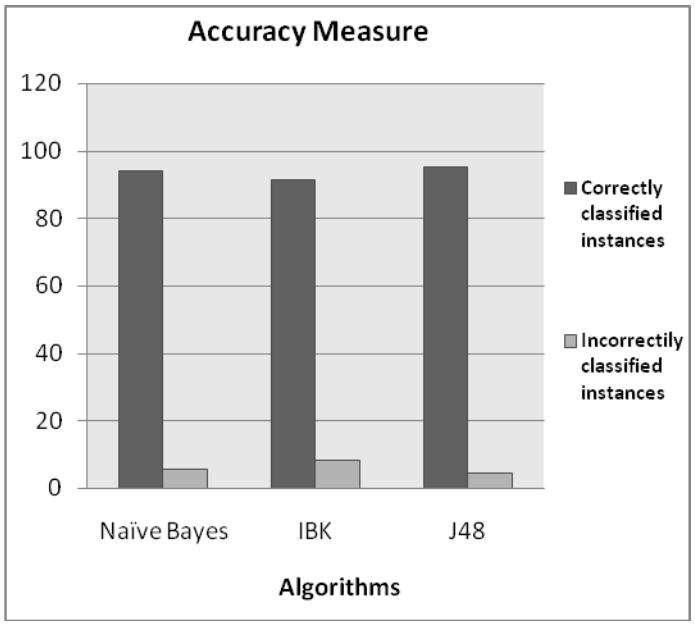

Fig 2: Accuracy Measure for given three Classifiers.

The time taken to build the classification model by using 10 fold cross validation are displayed in the Table 2.

Algorithm	**Time accuracy**
Naive Bayes	0.01 seconds
IBK	0 seconds
J48	0.05 seconds

Table 2: Time Accuracy for given three Classifiers

From the experimental results, it is inferred that the time accuracy for the J48 algorithm is higher than the Naive bayes and IBK algorithms for Breast Tissue dataset. The time accuracy for given classifiers are illustrated in Fig 3.

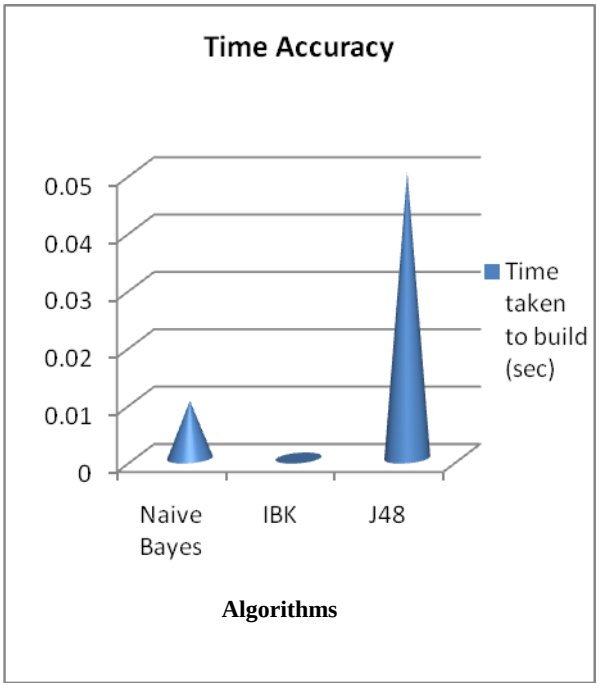

Fig 3: Time accuracy for given three Classifiers

The following table calculates the error measure for the three classifiers using the algorithms Naive bayes, IBK and J48. From the table, it is inferred that the cross validation parameter for the J48 algorithm, the MAE, RMSE, RAE and RRSE are lower than Naive bayes and IBK classification algorithms for breast tissue dataset. The Error rates measure for given three classifiers is illustrated in Fig 4.

Algorithm	MAE	RMSE	RAE	RRSE
Naive Bayes	0.0289	0.1423	10.436	38.2287
IBK	0.0431	0.1647	15.5682	44.2504
J48	0.0157	0.1254	5.6838	33.6979

Table 3: Error Measure for Three Classifiers

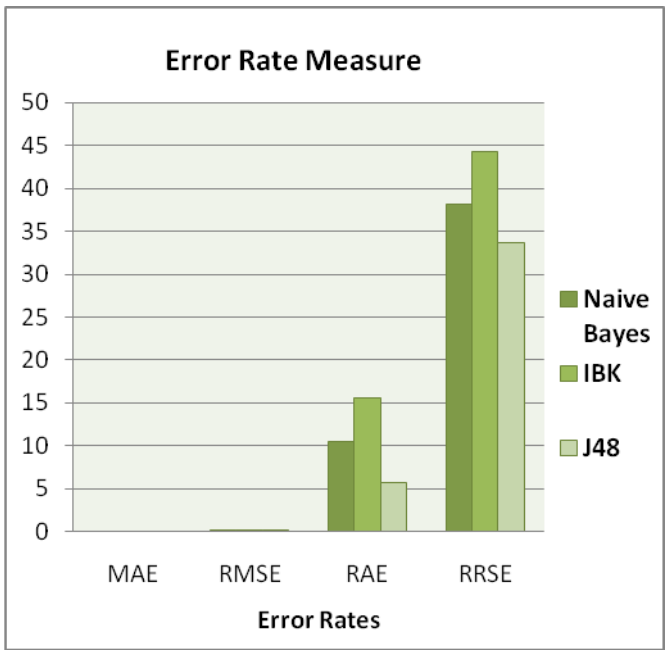

Fig 4: Error Measure for Three Classifiers

V. Conclusion

In this research we used the training dataset of breast tissue to classify the tissues using 10 fold cross validation and we have used three different classification algorithms to know the best classifier. The algorithm which has higher accuracy and the lowest mean absolute error is chosen as the best algorithm. By analysing the algorithms we concluded that the J48 algorithm has given better results than Naive bayes and IBK.

References

[1] Ahamed Lebbe Sayeth Saabith, Elankovan Sundararajan, Azuraliza Abu Bakar –"COMPARATIVE STUDY ON DIFFERENTCLASSIFICATION TECHNIQUES FOR BREAST CANCER DATASET" International Journal of Computer Science and Mobile Computing, Vol.3 Issue.10, October- 2014.

[2] American Cancer Society (2013).

[3] Gopala Krishna Murthy Nookala , Nagaraju Orsu ,Bharath Kumar Pottumuthu ,Suresh B. Mudunuri- "Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification" - (IJARAI) International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013.

[4] M.S. Chen, J. Han, and P.S. Yu. "Data mining: an overview from a database perspective," IEEETransactions on Knowledge and Data Engineering, Vol. 8, No.6, pp. 866 – 883, 2002.

[5] Dursun Delen, Glenn Walker, Amit Kadam "Predicting breast cancer survivability: A comparison of three data mining methods" Artificial Intelligence in Medicine (2004).

[6] K.R.Lakshmi, M.Veera Krishna, S.Prem Kumar-"performance comparison of data mining techniques for prediction and diagnosis of breast cancer disease survivability"- Asian journal of computer science and information technology.

[7] A.Priyanga, Dr.S.Prakasam "The Role of Data Mining-Based Cancer Prediction system (DMBCPS) in Cancer Awareness" International Journal of Computer Science and Engineering Communications- IJCSEC. Vol.1 Issue.1, December 2013.

[8] Vikas Chaurasia, Saurabh Pal "A novel approach for breast cancer detection using data mining techniques "International journal of innovative research in computer and communication engineering vol. 2, issue 1, January 2014.

[9] G.RaviKumar,Dr.G.A.Ramachandra,K.Nagamani "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques" International Journal of Innovations in Engineering and Technology (IJIET)

[10] Miss Jahanvi Joshi ,Mr. RinalDoshiDr. Jigar Patel-"Diagnosis and prognosis breast cancer using classification rules"- International Journal of Engineering Research and General Science Volume 2, Issue 6, October-November, 2014 ISSN 2091-2730.

[11] S. Aruna, Dr.S.P. Rajagopalan and L.V. Nandakishore "An empirical comparison of supervised learning algorithms in disease detection", IJITCS, August 2011.

[12] A. Endo, T. Shibata and H. Tanaka (2008), "Comparison of seven algorithms to predict breast cancer survival", Biomedical Soft Computing and Human Sciences, vol.13, pp.11-16.

[13] Dr. S.Vijayarani, S. Sudha, ―An Effective Classification Rule Technique for Heart Disease Prediction‖2010.

[14] https://archive.ics.uci.edu/ml/datasets/Breast+ Tissue

[15] http://www.cancer.org/healthy/findcancerear ly/womenshealth/non-cancerousbreastcondition s/non-cancerous-breast-conditions-normal-breast-tissue

[16] http://software.ucv.ro/~cmihaescu/ro/teaching /AIR/docs/Lab4-NaiveBayes.pdf

[17] http://en.wikipedia.org/wiki/K-nearest_neigh bors_algorithm

[18] http://www.opentox.org/dev/documentation /components/j48

[19] http://en.wikipedia.org/wiki/C4.5_algorithm