IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 15, Issue 5 (Nov. - Dec. 2013), PP 39-44
www.iosrjournals.org

Performance and Evaluation of Data Mining Techniques in Cancer Diagnosis

R.M. Chandrasekar Ph.D, V. Palaniammal M.C.A., M.Phil.
(Professor, Computer Science, Annamalai University, Chidambaram, Tamil Nadu, India.)
(Asst. Professor, Thiruvalluvar University College of Arts & Science, Kallakurichi, Tamil Nadu, India.)

Abstract: *We analyze the breast Cancer data available from the WBC, WDBC from UCI machine learning with the aim of developing accurate prediction models for breast cancer using data mining techniques. Data mining has, for good reason, recently attracted a lot of attention, it is a new Technology, tackling new problem, with great potential for valuable commercial and scientific discoveries. The experiments are conducted in WEKA. Several data mining classification techniques were used on the proposed data. There are many classification techniques in data mining such as Decision Tree, Rules NNge, Tree random forest, Random Tree, lazy IBK. The aim of this paper is to investigate the performance of different classification techniques. The data breast cancer data with a total 286 rows and 10 columns will be used to test and justify the different between the classification methods and algorithm.*

Keywords - *Machine learning, data mining Weka, classification, breast cancer*

I. Introduction

About 1 in 8 U.S women (just under 12%) will develop invasive breast cancer over the course of her life. In 2011, an estimated 230,480 new cases of invasive breast cancer were expected to be diagnosed in women in the U.S, along with 57,650 new cases of non-invasive (in situ) breast cancer. For women in the U.S breast cancer death rates are higher than those for any other cancer besides lung cancer. In 2011 there were more than 2.6million breast cancer survivor in the U.S.A women's risk of breast cancer approximately relative (mother, sister, daughter) who has been diagnosed with breast cancer. About 15% of women who get breast cancer have a family member diagnosed with it. About 85% of breast cancers occur in women who have no family history of breast cancer. These occur due to genetic mutations that happen as a result of the aging process and life in general.

Breast cancer (An overview):

Breast Cancer is the most common cancer diseases among women excluding no melanoma skin. Cancers are divided into two types benign and Malignant. If the cancer is benign under the conditions of early diagnosis. Malignancy status includes the three basic measurement 1.Age 2.Longer tumor length 3.ADC or Apparent Diffusion coefficient (biopsy confirmed).Attributes 1.patient id 2.diagnosis m=malignant ,b=benign 3.Real valued features. Cell uncles' 1.radius 2.texture 3.perimeter 4.Area 5.Smoothness.I Diagnosis Attributes are 1.Menopausal status a. Premenopausal, b.post menopausal. II.Basic diagnosis :-1.clinical examinations,2.mammography (y/n),3.MRI (y/n),4.Ultrasound (y/n),5.Fine Needle aspiration (y/n),6.Core biopsy (y/n),7.Open biopsy (y/n) ,8.Other (y/n),9.Unknown (y/n).III Clinical Trial Enrollment:-y/n/Not stated/inadequate. IV. Initial Representation:-Screening-mmaaography,Screening-MRI,Screening-Others,Symptomatic.V.Total extent of Lesion (DCIS AND INVASIVE) 1.Lesion in mm.VI.Lymphovascular Invasion(presence of tumor cells in endothelium-lined spaces) 1.present/absent/suspicious/not stated /unknown.

Number stages of Breast cancer:-

There are 4 number stages of breast cancer, staging takes into various factors, including 1.The size of the tumor (tumor means either breast lymph's or Area of cancer cells found on a Scan or mammogram) 2.Cancer cells have spread into nearby lymph glands (lymph Node) 3.The tumor cells has spread to any other part of the body (Metastasized-TNM). Stage1 breast cancer is split into 2 Stages.Stage1A tumor is 2 cm or smaller and has spread outside the breast.Stage1B: Small areas of breast Cancer cells are found in the lymph node closed to the breast and either the tumor is 2 cm or smaller. Stage 2 breast cancer: stage 2A:Tumor 2cm or smaller in the breast and cancer cells are found in 1 to 3 lymph Node in the armpit or in the lymph Node near the breast bone. Stage 2B: The tumor is larger than 2 cm but not larger than 5 cm and small areas of cancer cells are in the lymph Node.2 to 5 cm spread 1 to 3 lymph nodes in the armpit or near the breastbones or the tumor is larger than 5cm and has not spread to the lymph node. Stage 3A breast cancer No tumor is seen in the breast or the tumor may be any size and cancer is found in 4 to 9 lymph glands under the arm or in the lymph glands near the breast bone. The tumor is more than 5 cm and has spread into up to 3 lymph nodes near the breast bone.

Stage 3B: The tumor has spread to the skin of the breast or to the chest wall and made the kin break down or cancer swelling. The cancer may have spread to the up 9 lymph Node in the armpit or to the lymph glands near the breast bone. Stage 4 breast cancer: The tumor can be any size. The lymph Node may or may not contain cancer cells. The cancer has spread to the other parts of the body such as bone, lungs, liver, and brain. The TNM (Tumor, Node, and Metastasis) system is specifies for each type of cancer. Once a patient‘s T, N, and M categories have been determined, this information is combined in a process called stage grouping to determine a women disease stage.Stage0 (the least advantage stage) to Stage4 (the most advantage stage).

II. Headings

Existing System:

The data sets used are SEER data or Wisconsin data. Data pre-processing is applied before data mining to improve the quality of the data. Data pre-processing includes data cleaning, data integration, data transformation and data reduction techniques. The features used for classification purposes coincided with the Breast Imaging Reporting and Data System (BI-RADS) as this is how radiologists classify breast cancer. The BI-RADS features of density, mass shape, mass margin and abnormality assessment rank are used as they have been proven to provide good classification accuracy. A Classification method, Decision tree algorithms are widely used in medical field to classify the medical data for diagnosis. Feature Selection increases the accuracy of the Classifier because it eliminates irrelevant attributes. Feature selection with decision tree classification greatly enhances the quality of the data in medical diagnosis. CART algorithm with various feature selection methods to find out whether the same feature selection method may lead to best accuracy on various datasets of same domain. Artificial neural networks (ANNs) and support vector machines have been recently proposed as a very effective method for pattern recognition, machine learning and data mining. The discrimination capability of the features extracted from the sonograms was tested by using the SVM (support vector machine), ANN and KNN (K-nearest neighbor) classifier. It was found that the SVM gave the greatest accuracy while the ANN had the highest sensitivity. Back propagation neural network (BPNN) and radial basis function network (RBFN) were used for the training and testing of data.

Algorithm:

1. CART Algorithm
2. Decision tree algorithms
3. ID3
4. C4.5 decision tree algorithms
5. Hunt’s algorithm
6. SVM
7. Naïve Bayes classifier.
8. The back-propagated neural network.

Proposed System:

Early diagnosis needs an accurate and reliable diagnosis procedure that can be used by physicians to distinguish benign breast tumors from malignant ones without going for surgical biopsy. The objective of these predictions is to assign patients to one of the two group either a *“benign*” that is noncancerous or a *“malignant*” that is cancerous. The prognosis problem is the long-term care for the disease for patients whose cancer has been surgically removed. In this paper, we propose a model-based data mining technique with a neural network classification technique and the improvements possible using an ensemble approach. Ensembles have been proposed as a mechanism for improving the classification accuracy of existing classifiers providing that constituents are diverse. Contrasting the cluster analysis technique against a baseline neural network classifier, and then considers the effects of applying an ensemble technique to improve the accuracies. In different cases the state of the disease condition itself can be marked by stages where the diagnostic symptoms or signs can be subtle or different to other stages of the disease. This means that there is often not a clean mapping between the diagnostic features and the diagnosis. The usage of clustering has also been extended to classifiers and detection systems in order to improve detection and provide greater classification accuracy. Use of a feature selection technique with a decision tree classifier to classify breast cancer. Another advantage of the model-based clustering approach is that no decisions have to be made about the scaling of the observed variables: for instance, when working with normal distributions with unknown variances, the results will be the same irrespective of whether the variables are normalized or not. In order to fasten the classification we are using Artificial Neural Network (ANN).

Algorithm:

1. Decision tree algorithms – For feature selection

2. C4.5 Algorithm – For classification
3. ANN (Artificial Neural Network)

Software Requirement:

- The WEKA is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. The toolkit is developed in Java and is an open source software issued under the GNU General Public License.
- WBCD dataset.

III. Figures And Tables

Characteristics of database:

Wisconsin diagnostics Breast Cancer this database is created by William H.Wolberge at University of Wisconsin. This database contains 286 observations among which 201 are benign cases and 85 are malignant cases. These instances are described by 10 Attributes:-.class no-recurrence events, recurrence events, 2.age, 3.Menopause, 4.tumor_size, 5.inv_nodes, 6.node_cap, 7.deg_malig, 8.breast, 9.breast_quad, 10.irradiat

Table1:

Attributes	Values
Age	10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99
menopause premeno	lt40, ge40,
tumor-size	0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40- 44,45-49, 50-54,55-59
inv-nodes	0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26,27-29, 30-32, 33-35,36-39
node-caps	yes, no
deg-malig	1, 2, 3
Breast	left, right
breast-quad	left-up, left-low, right-up, right-low, central
irradiat	yes, no
Class	no-recurrence-events, recurrence-events

Classification prediction process in data mining:

Data mining is collection of techniques for efficient automated discovery of previously unknown valid novel, useful and understandable pattern in large databases. Classification is the process of minding a model that describes and distinguishes data classes concepts. This derived model is based on the analysis of a set of training data. Classification predicts categorial labels predictable numerical data values rather than class lables.

We used several classification methods resulting in identification of top 5 classification alogorithm. This paper presents a brief description of the classifiers and meta_classifiers used in the experiments results.

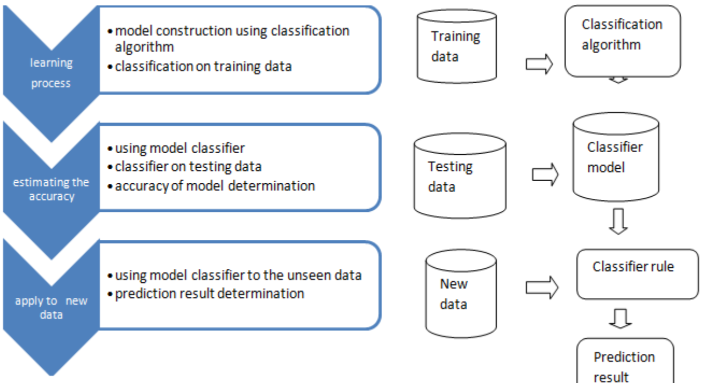

learning process
• model construction using classification algorithm
• classification on training data
estimating the accuracy
• using model classifier
• classifier on testing data
• accuracy of model determination
apply to new data
• using model classifier to the unseen data
• prediction result determination
Training data
Classification algorithm
Testing data
Classifier model
New data
Classifier rule
Prediction result

Fig1: Classification prediction process in data mining

Rules NNge (Non_Nested Generalised Examplars):

NNge Classifier new examples by determining the nearest neighbour in the exemplar database using a Euclidean distance function.

The function is $D_{EH} = W_H \sqrt{\sum_{I=1}^{M}\left(W_I \frac{E_{I-H_I}}{MAX_{I-MIN_I}}\right)^2}$

Lazy K Star classifier

Lasy kstar algorithm most successful classifier. Implementation of an instance based classifier, which uses the entropic distance measures. Test instance is based upon the class of those training instance.

Lazy IBK (K_Nearest Neighbours classifier):

IBK classifier is simple instance based learner that uses the class of nearest K training instances for the class of the test instance. Predicts the class of the single nearest training instance for each test instance. Get maximum no of instances allowed in the training pool.

Meta Decorate (Diverse Ensemble creation by oppositional relabeling of artificial Training Examples):

Meta learners such as Boosting Bagging and Random Forest provide diversity by sub sampling or reweighting the existing training data. Decorate performs by adding randomly constructed to the training set,when building new members.

$$d_{j(x)} = \begin{cases} 0 \; if \, c_{j(x)} \\ 1 \quad otherwise \end{cases} = c^*(x)$$

$$D = \frac{1}{NM}\sum_{j=1}^{m}\sum_{i=1}^{n} d_j(x_i)$$

$c^*(x)$ =arg max $p_{k(x)}$,kϵ{1…….Q)

$$p_{k(x)} = \sum_{c_{j\epsilon c^*}} pc_{3,k(x)} / \quad c^*$$

Tree Random Forest:

Random Forest classifier consists of multiple decision trees. The final class of an instance in a Random forest is assigned by outputting the class that is the mode of the outputs of individual trees. Which can produce robust and accurate classification. Random classifiers is based on a combination of many decision Tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. RF has excellent accuracy among current classifier Algorithm.

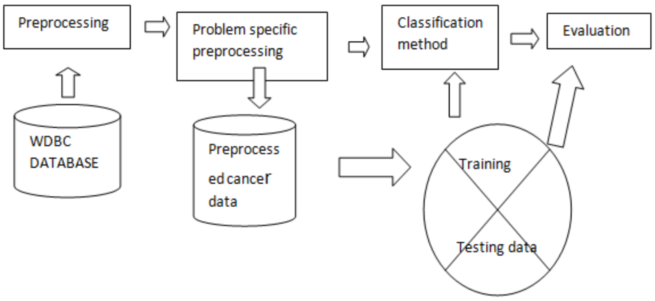


Fig2:Block diagram of the survival detection system

Table2:

Simulation result of each Algorithm

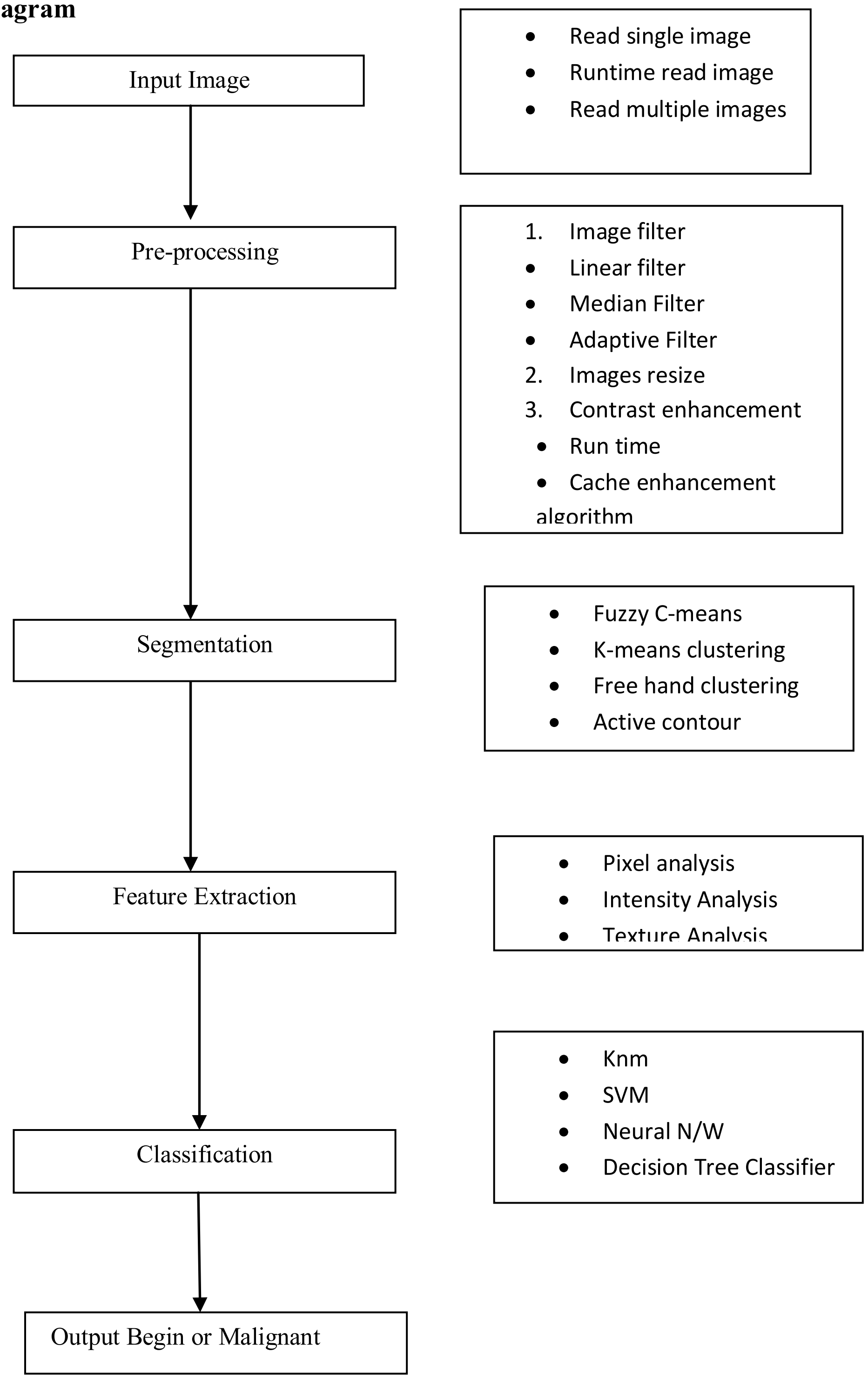
Algorithm **Total instance 286**	**Correctly classify** **Instance(%value)**	**Incorrectly classify** **Instance(%value)**	**Time taken** **(seconds)**	**Kuppa status**
Rule NNge	249 97.2028	8 2.7972	0.16	0.933
Random Forest	280 97.9021	6 2.0979	0.11	0.9501
Meta Decorate	259 90.5594	27 9.4406	1.81	0.7584
Lazy IBK	280 97.9021	6 2.0979	0	0.9491
Lazy Kstar	280 97.9021	6 2.0979	0	0.9491
Random Tree	280 97.9021	6 2.0979	0.02	0.9491

Table3:

Training and Simulation errors:

Algorithm Total instance 286	Mean Absolute Error %	Root Mean Squared Error %	Relative absolute Error%	Root Relative Squared Error%
Rules NNge	0.028	0.1672	6.6868	36.5947
Random Forest	0.0978	01614	23.3712	35.3247
Meta Decorate	0.3655	0.3797	87.3643	83.083
Lazy-IBK	0.0253	0.1053	6.0487	23.0348
Lazy-Kstar	0.0747	0.1399	17.8681	30.6094
Random Tree	0.0221	0.1052	5.2937	23.0237

Architecture Diagram



IV. Conclusion

This paper presents effective classification Techniques. After investigation of different classification Algorithm we have chosen 6 classifier based on our simulation performance and we have used Tree Random classifier achieved overall classification accuracy 98%, which is significanant. In future work we propose to analyze Ensemble classifier for 100% accuracy.

Acknowledgements

The authors like to express our appreciations to prof. R.M. Chandrasekar for his help in correcting earlier versions of this paper. We also would like to thank the anonymous reviewers for their valuable comments and in sightful suggestions.

References

[1] Mitchell, T.M. 1997. Machine Learning. McGraw-Hill Science

[2] John, G., Cleary, E. & Leonard, E. 1995. K*: An Instance-based Learner Using an Entropic Distance Measure. In: 12th International Conference on Machine Learning, 108-114.

[3] Aha, D. & Kibler, D. 1991. Instance-based learning algorithms. Machine Learning.

[4] American Cancer Societies. Breast Cancer Facts& Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[5] National Resource. Cancer Epidemiology Biomarkers & Prevention 1999; 8:1117-1121.

[6] Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.

[7] Lundin M, Lundin J, Burke HB, Toikkanen S Pylkkanen L, Joensuu H. Artificial neural Networks applied to survival prediction in breast cancer. Oncology 1999; 57:281-6.

[8] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data Mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[9] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Francisco: Morgan Kaufmann; 2005.

[10] Weka: Data Mining Software in Java,http://www.cs.waikato.ac.nz/ml/weka

[11] lando Anunciac¸ ˜ao and Bruno C. Gomes and Susana Vinga and Jorge Gaspar and Arlindo L. Oliveira and Jos´e Rueff , A Data Mining Approach for the detection of High-Risk Breast Cancer Groups

[12] aria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coma, .Application of Data Mining Techniques for Medical Image Classification.Proceeding of second International worshop on Mutimedia data mining(MDM/KDD'2001),in conjuction with ACM SIGKDD conference.SAN FRANCISCO,USA,AUG 26,2001.