Evaluation of Attribute Selection Methods with Tree based Supervised Classification-A Case Study with Mammogram Images

M.Vasantha
Research Scholar
Mother Teresa Womens University,
Tamil Nadu, Chennai

Dr.V.Subbiah Bharathy
Dean,Academics
DMI Engineering College,
Chennai,TamilNadu

ABSTRACT
Attribute selection is generally considered as a challenging work in the development of image data mining oriented applications. Attribute subset selection is mainly an optimization problem, which involves searching the space of possible feature subsets to select the one that is optimal or nearly optimal with respect to the performance measures accuracy, complexity etc., of the application. This paper presents a comparative evaluation of several attribute selection methods based on the performance accuracy of different tree based supervised classification for mammogram images of MIAS database.

Keywords
Data mining, Attribute selection, Feature subsets, mammogram images.

1. INTRODUCTION
Dimensionality reduction plays an important technique used in data mining. Attribute selection is a popularly used dimensionality reduction technique, which has been the focus of research in machine learning and data mining and used in medical image mining and analysis. It helps to construct simple, comprehensive classification models with classification performance. Even though several models exist for feature reduction and selection process only few will be suitable for an environment of the application. Thus it is necessary to study the suitability for attribute selection methods for our mammogram data base.

2. RELATED WORK
Attribute selection is mainly used to identify most relevant features with respect to the performance measure used to evaluate the subset of features related to the criteria of interest[10]. Several methods of attribute selection have been proposed [1],[4][5]. Hall and Holmes [2] performed a benchmark comparison of several attribute selection methods only for two supervised classification learning schemes C4.5 and naive Bayes. Liu and Schumann in their study [6] discussed four attribute selection methods : ReleifF, Correlation Based , Consistency based and wrapper algorithms. Surndra and Haun in [7] illustrated how attribute subset selection bias the classification learning and stated that the bias may not cause negative impact in classification as much as expected in regression. In this paper a comparative accuracy analysis of several attributes methods was performed based on the performance of major classification algorithm for the mammogram images database.

3. DATABASE SOURCES
Breast cancer is one of the major causes for the increase in mortality among middle-aged women, especially in developed countries [3]. In India, the death toll due to the breast cancer is increasing at a rapid pace (Gajalakshmi et al., 2009). This warrants for early detection and diagnosis. The digital mammogram database has been maintained in hospitals and breast screening centers for further research. In recent years, the development of automated mammographic classification system has been involving the analysis of tumor's shape, size and texture features [1-4].So it is necessary to monitor the performance of the system. The digital mammogram images used for analysis are taken from the Mammogram Image Analysis Society (MIAS) an online database for mammograms available for research from UK. The MIAS Digital Mammogram Data base contains 322 images representing 161 mammogram pairs.

Mammograms are difficult to interpret, and a preprocessing phase of the image is necessary to improve the quality of the images and make the feature extraction phase more reliable. Background noise elimination is necessary to enhance the visibility and detestability of tumors such as malignant or benign In this paper we performed low pass filter to remove noises and applied histogram equalization method for contrast enhancement.

Actually MIAS contains only the images and attribute selection cannot be directly applied on the images. So we have to extract features from the image.

The set of features useful for mammogram tumor analysis are categorized in to intensity histogram features, spatial features, shape feature, Gray Level Co-occurrence Matrix (GLCM) features, demographic features etc. We extracted intensity histogram features and Gray Level Co-occurrence Matrix(GLCM) features for this study. The following features were included for the analysis: Autocorrelation(A1) Contrast(A2), Correlation(A3), Cluster Prominence(A4), ClusterShade(A5), Dissimilarity(A6), Energy(A7), Entropy1(A8), Homogeneity1(A9), Maximum probability(A10), Sum of squares(A11), Sum average(A12), Sum variance(A13), Sum entropy(A14), Difference variance(A15), Difference

entropy(A16), Information measure of correlation(A17), information measure of correlation1(A19), information measure of correlation2(A20),Inverse difference normalized (A21), information difference moment normalized (A22), Mean(A21) Variance(A22), Skewness(A23), Kurtosis(A24), Entropy2 (A25), Energy(A26).

4. WEKA EXPERIMENT EDITORS

To perform benchmark experiment we used WEKA [8] an open source java based machine –learning workbench that can be run on any computer that has a java run time environment installed. It brings together many machine learning algorithm and tools under a common frame work. WEKA has two primary modes: experiment mode and exploration mode .The exploration mode allows easy access to all of WEKA's data preprocessing, learning, preprocessing, attribute selection and data visualization modules in an environment that encourages initial exploration of data. The experiment mode allows larger –scale experiments to be run with results stored in a database for retrieval and analysis.

5. ATTRIBUTE SELECTION AND CATEGORISING SELECTION METHODS

Under the data exploration mode, we explored almost all attribute selection modules applicable for the data to collect optimal subset of attributes. For the mammogram data base, the feature selection is performed to find more relevant attributes for all possible combinations of attribute evaluators and search methods. The results are listed in table1.

Table 1 -Attribute selection methods are grouped based on selected attributes

Attribute evaluator	Search method	Selected methods	Group
CfsSubsetEvaluator	Best First	A1,A6,A12,A18, A19,A24,A25,A26	I
CfsSubsetEvaluator	Exhaustive search	All Attributes	0
CfsSubset Evaluator	Genetic search	A6,A14, A18,A19, A24,A25,A26	II
CfsSubsetEvaluator	Greedy stepwise	A1,A6,A12,A18,A19,A24,A25,A26	I
CfsSubset Evaluator	Linear Forward Search	A1,A6,A12,A18, A19,A24,A25,A26	I
ConsistencySubsetEvaluator	Best First	A1,A6,A18, A25,A28	IV
ConsistencySubsetEvaluator	Exhaustive search	All Attributes	0
ConsistencySubsetEvaluator	Genetic search	A1, A3, A6, A17, A18, A25	III
ConsistencySubsetEvaluator	Greedy stepwise	A18,A25,A28	V
ConsistencySubsetEvaluator	Linear Forward Search	A1,A6,A18, A25,A28	1V

6. EVALUATING ATTRIBUTE SELECTION METHODS

The attribute selection methods are to be evaluated based on the various accuracy measures of classification algorithms. First using all the attributes without pruning the classification accuracy measures of different tree based classification algorithms are evaluated and are listed in table 3. Classification accuracy measures such as correctly classified instances and Root Mean Square error are used to compare the strength of various attribute selection methods. The results are plotted in the Figure 1 correspond to the accuracy measures and correctly classified instances.

Table 2: (Group -0) -All Attributes

Classification Algorithm- Tree – Type	% of Correctly Classified instances	Root mean square error
J48	80	0.3222
J48 graft	81.6667	0.3340
Simple CART	73.333	0.3888
Random Forest	76.6667	0.3145
Random tree	83.3337	0.3333
REP tree	70	0.4104

Table3:(Group-I)-Attributes- A1,A6,A12,A18,A19,A24,A25,A26

Classification Algorithm Tree type	% of Correctly Classified instances	Root mean square error
J48	83.3333	0.3209
J48 graft	83.3333	0.3209
Simple CART	75	0.3839
Random Forest	88.333	0.2576
Random tree	85	0.3162
REP tree	78.3333	0.3548

Table 4: (Group II)-Attributes-
A6, A14, A18, A19, A24, A25, A26

Classification Algorithm- Tree –Type	% of Correctly Classified instances	Root mean square error
J48	83.3333	0.3209
J48 graft	83.3333	0.3209
Simple CART	76.6667	0.3741
Random Forest	86.6667	0.2552
Random tree	90	0.2582
REP tree	80	0.3436

Table 5: (Group III) - Attributes - (A1, A3, A6, A17, A18, A25)

Classification Algorithm- Tree –Type	% of Correctly Classified instances	Root mean square error
J48	76	0.3636
J48 graft	78.3333	0.3479
Simple CART	68.3333	0.4286
Random Forest	78.3333	0.3195
Random tree	80	0.3651
REP tree	70	0.3891

Table 6: (Group IV) -Attributes - A1,A6,A18,A25,A28

Classification Algorithm- Tree –Type	% of Correctly Classified instances	Root mean square error
J48	80	0.3168
J48 graft	73.333	0.3497
Simple CART	78.3333	0.3758
Random Forest	80	0.2848
Random tree	83.333	0.3333
REP tree	68.888	0.3992

Table 7: (Group V) - Attributes - A18,A25,A28

Classification Algorithm- Tree –Type	% of Correctly Classified instances	Root mean square error
J48	75	0.3635
J48 graft	75	0.3635
Simple CART	75	0.3703
Random Forest	73.3333	0.3383
Random tree	78.33	0.3651
REP tree	70	0.3891

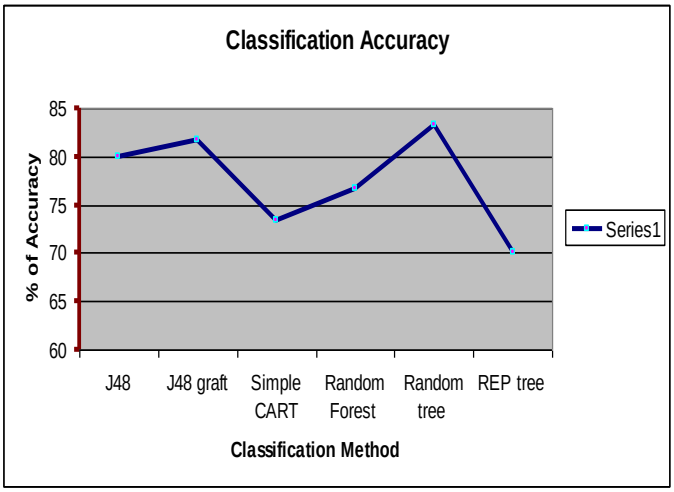

Figure 1: Classification Accuracy with All Features.

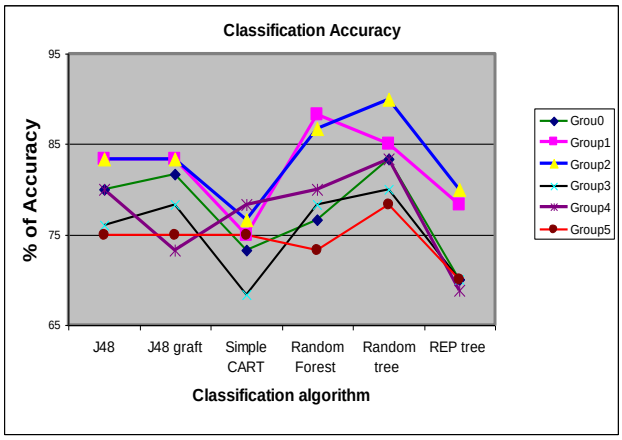

Figure 2: Classification Accuracy with selected features under Group0, Group I, GroupII, Group III ,Group IV and Group V

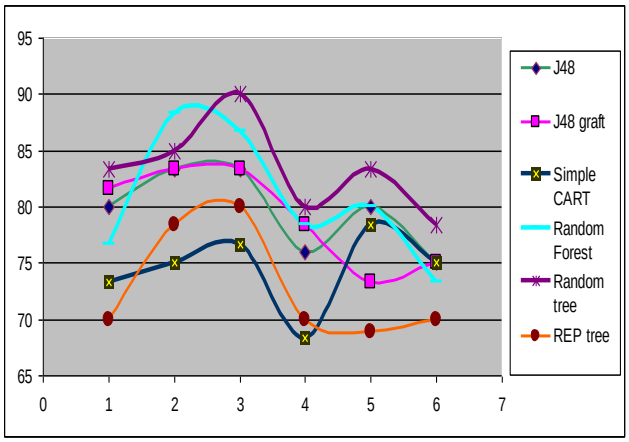

Figure 3: Classification Accuracy of different tree based classification Algorithms.

7. RESULTS AND CONCLUSION

From the tables 3 to 7 for the above selected classification algorithm, the comparison of different groups of attribute selection algorithm is carried out. Based on the percentage of correctly classified instances, the result is shown in the Figure2. From the figure, the group I and group II type of attribute selection is performing better classification. From the Figure 3, Random Tree and Random forest are performing well. For comparative analysis we have used single data base consisting of 60 images. Further the hybrid and integrated type of feature selection algorithms can be selected and evaluated.

8. REFERENCES

[1] M.Dash and H.Liu, Feature selection for Classification, In Intelligent Data Analysis, vol. 1, no. 3,1997

[2] M.A.Hall and G.Homes, Benchmarking attribute selection Techniques for Discrete Class Data Mining, IEEE Transactions on Knowledge and Data engineering, vol.15, Issue 6 , Nov-Dec. 2003 ,pp. 1437-1447.

[3] R. G. Bird, R. G. Wallace, and B. C. Yankaskas, "Analysis of cancers missed at Screening mammography *Radiology*, vol. 184, pp. 613–617, 1992.

[4] G.John, R.Kohavi and K.Pfleger,Irrelevent features and subset selection problem, In Morgan Kaufmann New Brunswick,NJ, editor, The eleventh International conference on Mobile Learning, 1994,pp-121-129.

[5] H.Liu and H.Motoda ,Feature selection for knowledge discovery and data mining ,In Kluver International Series in Engineering and computer Science Secs,1998.

[6] Y.liu and M.Schumaan , Data mining feature selection for credit scoring models , Journal of Operation Research Society (2005) 56, pp. 1099-1108,published online 20 April 2005.

[7] K. Surendhra and Singhi,Huan Liu ,Feature Subset selection Bias for Classification Learning :Appearing in proceedings of 23 rd International Conference on Machine Learning, Pitsburgh,PA,2006..

[8] WEKA: Waikato environment for knowledge analysis .http://www.cs.waikato.ac. nz/ml/weka