

ANALYSIS OF FEATURE SELECTION WITH CLASSIFICATION: BREAST CANCER DATASETS

D.Lavanya *

Department of Computer Science, Sri Padmavathi Mahila University
Tirupati, Andhra Pradesh, 517501, India
lav_dlr@yahoo.com
<http://www.spmvv.ac.in>

Dr.K.Usha Rani

Department of Computer Science, Sri Padmavathi Mahila University
Tirupati, Andhra Pradesh, 517501, India
usharanikurubar@yahoo.co.in
<http://www.spmvv.ac.in>

Abstract

Classification, a data mining task is an effective method to classify the data in the process of Knowledge Data Discovery. 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. This paper analyzes the performance of Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets. The results show that a particular feature selection using CART has enhanced the classification accuracy of a particular dataset.

Keywords: Data Mining; Feature Selection Classification; Decision Tree; CART; Breast Cancer Datasets.

1. Introduction

Knowledge Data Discovery (KDD) is a process of deriving hidden knowledge from databases. KDD consists of several phases like Data cleaning, Data integration, Data selection, Data transformation, Data mining, Pattern evaluation, Knowledge representation. Data mining is one of the important phases of knowledge data discovery. Data mining is a technique which is used to find new, hidden and useful patterns of knowledge from large databases. There are several data mining functions such as Concept descriptions, Association Rules, Classification, Prediction, Clustering and Sequence discovery to find the useful patterns.

Data preprocessing is applied before data mining to improve the quality of the data. Data preprocessing includes data cleaning, data integration, data transformation and data reduction techniques. Cleaning is used to remove noisy data and missing values. Integration is used to extract data from multiple sources and storing as a single repository. Transformation transforms and normalizes the data in a consolidated form suitable for mining. Reduction reduces the data by adopting various techniques i.e., aggregating the data, attribute subset selection, dimensionality reduction, numerosity reduction and generation of concept hierarchies. The data reduction is also called as feature selection. Feature selection or attribute selection identifies the relevant attributes which are useful to the data mining task. Applying feature selection with data mining technique improves the quality of the data by removing irrelevant attributes.

Classification is extensively used in various application domains: retail target marketing, fraud detection, design of telecommunication service plans, Medical diagnosis, etc [Brachman.R et al., 1996], [K.U.M Fayyad et al.,1996]. In the domain of medical diagnosis classification plays an important role. Since large volume of data maintained in the medical field, classification is extensively used to make decisions for diagnosis and prognosis of patient's disease. Decision tree classifiers are used extensively for diagnosis of diseases such as breast cancer, ovarian cancer and heart sound diagnosis and so on [Antonia Vlahou et al.,2003], [Kuowj et al.,2001], [Stasis

A.C et al.,2003]. Feature selection with decision tree classification greatly enhances the quality of the data in medical diagnosis.

In this study we considered three breast cancer datasets for experimental purpose to analyze the performance of decision tree classifier, 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. The paper has been organized in six sections. Section 2 describes the related work to this study. Section 3 deals the fundamental concepts of classification and decision trees. In Section 4 Feature Selection mechanisms are presented. The experimental results are presented with explanation in Section 5 followed by conclusions in Section 6.

2. Related Work

Classification is one of the most fundamental and important task in data mining and machine learning. Many of the researchers performed experiments on medical datasets using decision tree classifier. Few are summarized here:

In 2010, Asha Gowda Karegowda et al.[Asha Gowda Karegowda et al.,2010] proposed a wrapper approach with genetic algorithm for generation of subset of attributes with different classifiers such as C4.5,Naïve Bayes, Bayes Networks and Radial basis functions. The above classifiers are experimented on the datasets Diabetes, Breast cancer, Heart Statlog and Wisconsin Breast cancer.

About Ella Hassaneian [Hassaneian, 2003] in 2003 had experimented on breast cancer data using feature selection technique to obtain reduced number of relevant attributes, further decision tree-ID3 algorithm is used to classify the data.

In 2005, Kemal Polat et al. [Kemal Polat et al., 2005], proposed a new classification algorithm feature selection-Artificial Immune Recognition System (FS-AIRS) on breast cancer data set. To reduce the data set C4.5 decision tree algorithm is used as a feature selection method.

Deisy.C et al. in 2007 [Deisy. C et al., 2007] experimented breast cancer data using three feature selection methods Fast correlation based feature selection, Multi thread based FCBF feature selection and Decision dependent-decision independent correlation further the data is classified using C4.5 decision tree algorithm.

Mark A. Hall et al. [Mark A. Hall et.al., 1997] in 1997 have done experiments on various data sets using Correlation based filter feature selection approach further the reduced data is classified using C4.5 decision tree algorithm.

In 2011, D. Lavanya et al. [D.Lavanya et al., 2011] analyzed the performance of decision tree classifiers on various medical datasets in terms of accuracy and time complexity.

3. Classification

Classification [J.Han et al.,2000], a data mining task which assigns an object to one of several pre-defined categories based on the attributes of the object. The input to the problem is a data-set called the training set, which consists of a number of examples each having a number of attributes. The attributes are either continuous, when the attribute values are ordered, or categorical when the attribute values are unordered. One of the categorical attributes is called the class label or the classifying attribute. The objective is to use the training set to build a model of the class label based on the other attributes such that the model can be used to classify new data not from the training data-set. Classification has been studied extensively in statistics, machine learning, neural networks and expert systems over decades [Mitchell, 1997]. There are several classification methods:

- Decision tree algorithms
- Bayesian algorithms
- Rule based algorithms
- Neural networks
- Support vector machines
- Associative classification
- Distance based methods
- Genetic Algorithms

3.1 Decision Trees

Decision tree induction [J.Han et al., 2000] is a very popular and practical approach for pattern classification. Decision tree is constructed generally in a greedy, top down recursive manner. The tree can be constructed in a breadth first manner or depth first manner. Decision tree structure consists of a root node, internal nodes and leaf nodes. The classification rules are derived from the decision tree in the form of - if then else. These rules are used to classify the records with unknown value for class label. The decision tree is constructed in two phases: Building Phase and Pruning Phase.

In Building phase of the tree the best attributes are selected based on attribute selection measures such as Information gain, Gain Ratio, Gini Index, etc. Once the best attribute is selected then the tree is constructed with that node as the root node and the distinct values of the attribute are denoted as branches. The process of selecting best attribute and representing the distinct values as branches are repeated until all the instances in the training set belong to the same class label.

In Pruning phase the sub trees are eliminated which may over fit the data. This enhances the accuracy of a classification tree. Decision trees handle continuous and discrete attributes. Decision trees are widely used because they provide human readable rules, easy to understand, construction of decision tree is fast and it yields better accuracy.

There are several algorithms to classify the data using decision trees. The frequently used decision tree algorithms are ID3, C4.5 and CART [Matthew N Anyanwu et al.,]. In this study the CART algorithm is chosen to classify the breast cancer data because it provides better accuracy for medical data sets than ID3, C4.5 algorithms [D.Lavanya et al., 2011]. CART [Breiman et al., 1984] stands for Classification and Regression Trees introduced by Breiman. It is based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It also handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3 [Quinlan, 1986], C4.5 [Quinlan, 1992]. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

4. Feature Selection

Feature selection (FS) plays an important role in classification. This is one of the Preprocessing techniques in data mining. Feature selection is extensively used in the fields of statistics, pattern recognition and medical domain. Feature Selection means reducing the number of attributes. The attributes are reduced by removing irrelevant and redundant attributes, which do not have significance in classification task. The feature selection improves the performance of the classification techniques. The process of feature selection is

- Generation of candidate subsets of attributes from original feature set using searching techniques.
- Evaluation of each candidate subset to determine the relevancy towards the classification task using measures such as distance, dependency, information, consistency, classifier error rate.
- Termination condition to determine the relevant subset or optimal feature subset.
- Validation to check the selected feature subset.

The feature selection process [Mark A.Hall et al., 1997] is represented in figure 1.

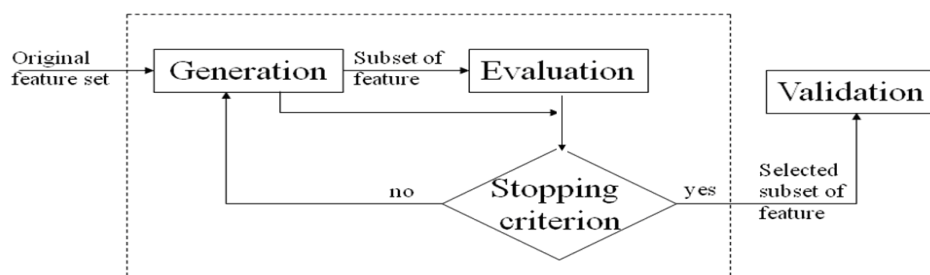


Fig. 1. Feature Selection Process

Feature selection methods are classified as filter, wrapper and hybrid approaches. The filter approach applies to the data before the classification. In this method features are evaluated by using heuristics based on general characteristics of the data. In the wrapper approach, the features are evaluated using the classification algorithms. In Hybrid approach features are evaluated using filter and wrapper approaches. Further, the reduced dataset is considered for classification.

5. Experimental Results

In this experiment the medical data related to breast cancer is considered because the breast cancer is one of the leading causes of death in women. The experiments are conducted using weka tool. In this study CART algorithm is chosen to analyze the breast cancer datasets because it provides better accuracy for medical data sets than the other two frequently used decision tree algorithms ID3 and C4.5 [D.Lavanya et al., 2011]. With an intension to find out whether the same feature selection method may lead to best accuracy for various datasets of same domain, various experiments are conducted on three different breast cancer datasets. Further to analyze the importance of feature selection in decision tree classifier-CART, we considered three different Breast Cancer Datasets: Breast Cancer, Breast Cancer Wisconsin (Original) and Breast Cancer Wisconsin (Diagnostic) with different attribute types. The data is collected from UCI machine learning repository [www.ics.uci.edu] which is publicly available. The description of the datasets is given in Table 1.

Table 1: Description of Breast Cancer Datasets

Dataset	No. of Attributes	No. of Instances	No. of Classes	Missing values
Breast Cancer	10	286	2	yes
Breast Cancer Wisconsin (Original)	11	699	2	yes
Breast Cancer Wisconsin (Diagnostic)	32	569	2	no

For consistency the missing values in the datasets are replaced with mean value of the respective attribute. The experiments are conducted on the above datasets with and without feature selection methods. And the results are compared and analyzed. The performance of the classifier is analyzed in terms of accuracy, time taken to execute to a model and tree size.

The performance of CART algorithm related to breast cancer datasets without feature selection is shown in table 2.

Table 2: CART algorithm – without feature selection

Datasets	Accuracy (%)	Time (Secs)	Tree Size
Breast Cancer	69.23	0.23	5
Breast Cancer Wisconsin (Original)	94.84	0.44	15
Breast Cancer Wisconsin (Diagnostic)	92.97	0.73	17

Further the experiments are conducted on 13 feature section methods. The supporting search techniques vary from one FS method to another. Hence, by considering the entire supporting search techniques related to a particular feature selection method various experiments are conducted and the results are shown in the tables 3 - 7. The search technique with best accuracy is highlighted for quick reference in those tables.

The highlighted best search technique corresponding to a particular feature selection method for all Breast Cancer Datasets is posted in the tables 8-10. By comparing all the feature selection methods the SVMAttributeEval method with accuracy of 73.03% is best for Breast Cancer Dataset, PrincipalComponentsAttributeEval method is the best one with accuracy of 96.99% for Breast Cancer Wisconsin (Original) Dataset and SymmetricUncertAttributesetEval method is best for Breast Cancer Wisconsin (Diagnostic) Dataset with 94.72% accuracy .

Table 3: FS Method - CfssSubsetEval.

Search Technique	Breast Cancer Dataset				Breast Cancer Wisconsin (Original)				Breast Cancer Wisconsin (Diagnostic)			
	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size
Best-First	5	70.27	0.14	5	9	94.84	0.17	15	12	92.97	0.3	15
Exhaustive	5	70.27	0.14	5	9	94.84	0.14	15	-	-	-	-
Genetic	5	70.27	0.14	5	9	94.84	0.14	15	14	93.49	0.44	5
Greedy-step wise	5	71.32	0.14	5	9	94.84	0.14	15	12	92.61	0.24	15
Linear forward selection	5	70.27	0.13	5	9	94.84	0.14	15	12	92.97	0.3	15
Random	6	70.27	0.28	5	9	94.84	0.14	15	-	-	-	-
Rank	5	70.97	0.13	5	9	94.84	0.14	15	11	93.32	0.3	15
Scatter	5	70.27	0.11	5	9	94.84	0.27	15	12	92.26	0.28	15
Subsize forward selection	5	71.32	0.19	5	9	94.84	0.24	15	12	92.79	0.58	15

Table 4: FS Method - ClassifierSubsetEval

Search Technique	Breast Cancer Dataset				Breast Cancer Wisconsin (Original)				Breast Cancer Wisconsin (Diagnostic)			
	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size
Best-First	6	67.83	6.25	29	6	93.99	3.31	19	7	93.32	48.34	27
Exhaustive	6	63.98	57.84	39	6	94.84	42.17	37	-	-	-	-
Genetic	6	67.48	8.58	29	6	93.41	10.45	37	9	94.05	87.31	25
Greedy-step wise	4	66.78	2.03	11	3	94.70	1.92	15	5	94.02	20.59	27
Linear forward selection	6	67.13	7.75	29	6	94.13	4.06	19	11	92.79	68.39	25
Race	5	70.62	19.61	7	3	94.99	5.2	9	7	92.99	50.19	15
Random	6	66.78	14.95	29	7	93.70	13.16	37	-	-	-	-
Rank	3	70.62	5.52	9	6	95.13	1.23	17	9	93.76	30.05	15
Scatter	4	70.27	5.52	9	4	94.42	5.94	33	11	92.97	28.78	25
Subset size forward selection	4	67.48	2.25	11	3	94.13	2.42	15	7	93.32	28.56	15

Table 5: FS Method - ConsistencySubsetEval

Search Technique	Breast Cancer Dataset				Breast Cancer Wisconsin (Original)				Breast Cancer Wisconsin (Diagnostic)			
	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size
Best-First	7	69.58	0.2	7	7	93.70	0.13	9	7	94.20	0.66	21
Exhaustive	7	69.23	0.61	7	6	95.13	0.41	9	-	-	-	-
Genetic	7	69.23	0.2	7	6	94.84	0.17	37	9	92.44	0.66	15
Greedy-step wise	7	69.23	0.17	7	7	93.99	0.16	9	7	94.20	0.27	21
Linear forward selection	8	70.27	0.2	7	7	93.84	0.14	9	7	93.14	0.36	23
Random	9	69.23	0.22	5	7	93.56	0.22	37	-	-	-	-
Rank	9	69.23	0.2	5	9	94.84	0.13	15	21	92.97	0.45	17
Scatter	9	70.62	0.44	5	7	94.56	0.2	19	10	94.55	2.8	13
Subset size forward selection	8	70.27	0.3	5	7	94.27	0.25	9	7	93.14	0.78	23

Table 6(Continued): FS Method - FilteredSubsetEval

Search Technique	Breast Cancer Dataset				Breast Cancer Wisconsin (Original)				Breast Cancer Wisconsin (Diagnostic)			
	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size
Best-First	5	70.27	0.13	5	9	94.84	0.11	15	9	92.61	0.25	7

Exhaustive	5	70.27	0.14	5	9	94.84	0.13	15	-	-	-	-
Genetic	5	71.32	0.14	5	9	94.84	0.13	15	12	93.49	0.28	19
Greedy-step wise	5	71.32	0.13	5	9	94.84	0.16	15	9	92.26	0.2	7
Linear forward selection	5	70.27	0.14	5	9	94.84	0.14	15	9	92.61	0.39	7
Random	6	70.27	0.14	5	9	94.84	0.14	15	-	-	-	-
Rank	5	70.27	0.16	5	9	94.84	0.13	15	7	92.97	0.39	5
Scatter	5	70.27	0.13	5	9	94.84	0.13	15	9	92.26	0.27	7
Subset size forward selection	5	71.32	0.19	5	9	94.84	0.24	15	9	92.44	0.24	7

Table 7: Other Feature Selection methods

Feature Selection Method	Search Technique	Breast Cancer Dataset				Breast Cancer Wisconsin (Original)				Breast Cancer Wisconsin (Diagnostic)			
		Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Tree Size	Reduced No of Attributes	Accuracy (%)	Time (Sec)	Size
ChiSquaredAttributeEval	Ranker	9	69.23	0.22	9	9	94.56	0.09	15	31	92.61	0.66	17
FilteredAttributeEval	Ranker	9	69.23	0.39	9	9	94.56	0.13	15	31	92.61	0.81	17
InfoGainAttributeEval	Ranker	9	69.23	0.22	9	9	94.56	0.13	15	31	92.61	0.47	17
GainRatioAttributeEval	Ranker	9	69.23	0.22	9	9	94.42	0.13	15	31	92.26	0.59	17
ReliefFAttributeEval	Ranker	9	69.23	0.3	9	9	94.56	0.83	15	31	92.79	1.78	17
PrincipalComponentsAttributeEval	Ranker	9	70.63	0.47	9	9	96.99	0.19	3	11	92.09	0.41	21
SVMAttributeEval	Ranker	9	73.07	19.3	9	9	94.56	1.41	15	9	94.56	1.41	15
SymmetricUncertAttributeEval	Ranker	9	69.23	0.22	9	9	94.42	0.14	15	31	92.26	0.72	17
SymmetricUncertAttributesetEval	FCBF	2	66.78	0.19	2	8	93.99	0.13	9	8	94.72	0.27	19

Table 8: Result of all feature selection methods - Breast Cancer Dataset

Feature selection Technique	Reduced No. of Attributes	Accuracy (%)	Time (Sec)	Tree size
CfssSubsetEval	5	71.32	0.14	5
ChiSquaredAttributeEval	9	69.23	0.22	5
ClassifierSubsetEval	6	95.13	1.23	17
ConsistencySubsetEval	9	70.62	0.44	5
FilteredAttributeEval	9	69.23	0.39	5
FilteredSubsetEval	5	71.32	0.13	5
GainRatioAttributeEval	9	69.23	0.22	5
InfoGainAttributeEval	9	69.23	0.22	5
ReliefFAttributeEval	9	69.23	0.3	5
PrincipalComponentsAttributeEval	9	70.63	0.47	5
SVMAttributeEval	9	73.07	19.3	5
SymmetricUncertAttributeEval	9	69.23	0.22	5
SymmetricUncertAttributesetEval	2	66.78	0.19	1

Table 9: Result of all feature selection methods - Breast Cancer Wisconsin (Original) Dataset.

Feature selection Technique	Reduced No. of attributes	Accuracy (%)	Time (Sec)	Tree size
CfssSubsetEval	9	94.84	0.14	15
ChiSquaredAttributeEval	9	94.56	0.09	15
ClassifierSubsetEval	6	95.13	1.23	17
ConsistencySubsetEval	6	95.13	0.41	9
FilteredAttributeEval	9	94.56	0.13	15
FilteredSubsetEval	9	94.84	0.11	15
GainRatioAttributeEval	9	94.42	0.13	15
InfoGainAttributeEval	9	94.56	0.13	15
ReliefFAttributeEval	9	94.56	0.83	15
PrincipalComponentsAttributeEval	9	96.99	0.19	3
SVMAttributeEval	9	94.56	1.41	15
SymmetricUncertAttributeEval	9	94.56	1.41	15
SymmetricUncertAttributesetEval	8	93.99	0.13	9

Table 10: Result of all feature selection methods - Breast Cancer Wisconsin (Diagnostic) Dataset

Feature selection Technique	Reduced No. of Attributes	Accuracy (%)	Time (Sec)	Tree size
CfssSubsetEval	14	93.49	0.44	5
ChiSquaredAttributeEval	31	92.61	0.66	17
ClassifierSubsetEval	9	94.05	87.31	25
ConsistencySubsetEval	10	94.55	2.8	13
FilteredAttributeEval	31	92.61	0.81	17
FilteredSubsetEval	12	93.49	0.28	19
GainRatioAttributeEval	31	92.26	0.59	17
InfoGainAttributeEval	31	92.61	0.47	17
ReliefFAttributeEval	31	92.79	1.78	17
PrincipalComponentsAttributeEval	11	92.09	0.41	21
SVMAttributeEval	9	94.56	1.41	15
SymmetricUncertAttributeEval	31	92.26	0.72	17
SymmetricUncertAttributesetEval	8	94.72	0.27	19

6. Conclusion

Accuracy is most important in the field of medical diagnosis to diagnose the patient's disease. Experimental results show that Feature Selection, a Preprocessing technique greatly enhances the accuracy of classification. We also conclude that the classifier accuracy has been surely enhanced by the use of any of Feature selection method than the classifier accuracy without feature selection. With an intension to find out whether the same feature selection method may lead to best accuracy for various datasets of same domain, various experiments are conducted on three different breast cancer datasets. The performance of Decision tree classifier-CART with and without feature selection in terms of accuracy, time to build a model and size of the tree on various Breast Cancer Datasets are observed. From the results it is clear that, though we considered only breast cancer datasets, a specific feature selection may not lead to the best accuracy for all Breast Cancer Datasets. The best feature selection method for a particular dataset depends on the number of attributes, attribute type and instances. Hence, whenever another dataset is considered, one has to experiment on that with various feature selection methods to identify the best one to enhance the classifier accuracy instead of simply considering the previously proved one related to the same domain. Once the best feature selection method is identified for a particular dataset the same can be used to enhance the classifier accuracy.

References

- [1] Matthew N.Anyanwu, Sajjan G.Shiva, "Comparative Analysis of Serial Decision Tree Classification Algorithms", International Journal of Computer Science and Security, volume 3.
- [2] R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky-Shapiro and E. Simoudis, "Mining Business Databases", Comm. ACM, Vol. 39, no. 11, pp. 42-48, 1996.
- [3] Breiman, Friedman, Olshen, and Stone. "Classification and Regression Trees", Wadsworth, 1984, Mezzovico, Switzerland.
- [4] Deisy.C, Subbulakshmi.B, Baskar S, Ramaraj.N, "Efficient Dimensionality Reduction Approaches for Feature Selection, Conference on Computational Intelligence and Multimedia Applications, 2007.
- [5] K U.M. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to knowledge Discovery in Databases", AI Magazine, vol 17, pp. 37-54, 1996.
- [6] Mark A. Hall, Lloyd A. Smith, Feature Subset Selection: A Correlation Based Filter Approach, In 1997 International Conference on Neural Information Processing and Intelligent Information Systems (1997), pp. 855-858.
- [7] Aboul Ella Hassaneian, "Classification and feature selection of breast cancer data based on decision tree algorithm, Studies and Informatics Control ,vol12, no1,March 2003.
- [8] J. Han and M. Kamber, "Data Mining; Concepts and Techniques, Morgan Kaufmann Publishers", 2000.

- [9] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath, *Feature Subset Selection Problem using Wrapper Approach in Supervised Learning*, International Journal of Computer Applications 1(7):13–17, February 2010.
- [10] Kuowj, Chang RF, Chen DR and Lee CC, " *Data Mining with decision trees for diagnosis of breast tumor in medical ultrasonic images*" ,March 2001.
- [11] D.Lavanya, Dr.K.Usha Rani, *Performance Evaluation of Decision Tree Classifiers on Medical Datasets*. International Journal of Computer Applications 26(4):1-4, July 2011.
- [12] T. Mitchell, "*Machine Learning*", McGraw Hill, 1997.
- [13] Kemal Polat, Seral Sahan, Halife Kodaz and Salih Günes, *A New Classification Method for Breast Cancer Diagnosis: Feature Selection Artificial Immune Recognition System (FS-AIRS)*, In Proceedings of ICNC (2)'2005. pp.830–838
- [14] J.R.Quinlan, "*c4.5: Programs for Machine Learning*", Morgan Kaufmann Publishers, Inc, 1992.
- [15] Quinlan, J.R, "*Induction of decision trees*". Journal of Machine Learning 1(1986) 81-106.
- [16] Stasis, A.C. Loukis, E.N. Pavlopoulos, S.A. Koutsouris, D. "*Using decision tree algorithms as a basis for a heart sound diagnosis decision support system*", Information Technology Applications in Biomedicine, 2003. 4th International IEEE EMBS Special Topic Conference, April 2003.
- [17] Antonia Vlahou, John O. Schorge, Betsy W.Gregory and Robert L. Coleman, "*Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data*", Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314.
- [18] www.ics.uci.edu/~mllearn/MLRepository.html