A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis

T.Sridevi
Research Scholar,
Mother Teresa Women's University, Kodaikanal,
Tamil Nadu,India

A.Murugan
Department of Computer Science,
Dr. Ambedkar Govt. Arts College, Chennai, Tamil Nadu, India

ABSTRACT
A major area of current research in data mining is the field of medical diagnosis. In the present study using the Breast cancer Wisconsin data sets, a feature selection algorithm Modified Correlation Rough Set Feature Selection (MCRSFS) predicts both diagnosis and prognosis by comparing several data mining classification algorithms. In the proposed approach, in level 1 of feature selection, features are selected based on rough set with different starting values of reduct. In level 2 features are selected from the reduced set based on the Correlation Feature Selection (CFS). Experiments show the proposed method is effective by comparing with others in terms of number of selected features and classification performance.

General Terms
Pattern Recognition, Machine learning.

Keywords
Data mining, feature selection, rough set, correlation, breast cancer.

1. INTRODUCTION
Data mining is the process of extracting useful and related information form a database [1]. Feature Selection (FS) is an important concept in pattern recognition and data mining. It aims to select the distinguishing features from a set of features and eliminating unnecessary features. Rough set theory can be used as a tool to reduce unnecessary features and to deal with vagueness and uncertainty in datasets. The main concept in rough set theory is to define the necessity of features. The measures of necessity are calculated by the functions of approximations. Rough set has been used to improve the efficiency and effectiveness of classification and applied for classification in various applications [2]. Breast cancer is considered as a major health issue for women. For the diagnosis of malignant ones in the Wisconsin Diagnostic Breast Cancer (WDBC) data set as well as for the recurrence of breast cancer in the Wisconsin Prognostic Breast Cancer (WPBC) data set, many techniques have been discussed [3], [4], [5] and [6].

In the present study both Breast cancer Wisconsin Diagnostic and Prognostic datasets are used. Two levels of feature reductions are proposed for breast cancer detection and prognosis problems. In level 1 of feature reduction, minimal feature subset is selected based on rough set and in level 2 features are selected from the minimal feature subset obtained from level 1 based on the correlation feature selection. Furthermore, different classification techniques are used to study the impact of the selected features.

2. MATERIALS AND METHODS
2.1Rough Set Feature Selection
Rough set theory was introduced by Pawlak in 1982 [7]. In rough set theory, an information table is defined as I = (U,C∪ D,V,f) where U is the universe of primitive objects and C a collection of condition attributes, D a collection of decision attributes, V a set of values of attributes in C and f:C→V a description function. For any P⊆C, there is an equivalence relation IND (P) as follows:

$$IND(P) = \{(x,y) \in U^2 \mid \forall a \in P\ a(x) = a(y)\}$$

If (x,y)∈ IND(P), then x and y are indiscernible by attributes from P. Assuming *P* and *Q* are equivalence relations in *U*, the important concept positive region$POS_P(Q)$ is defined as:

$$POS_P(Q) = \cup_{x \in U/Q} \underline{P}X$$

The positive region contains all objects of *U* that can be classified with certainty into classes of U/Q using attributes from P.

QuickReduct algorithm [8] is usually used to generate minimal feature subset. This algorithm uses the degree of dependency

$$\gamma_P(Q) = \frac{\| POS_P(Q) \|}{\| U \|}$$

It starts with an empty set of attributes. The best of the original attributes is determined and added to the set iteratively using the above dependency. The process is repeated until the dependency of the reduct candidate equals to 1.

2.2Correlation Feature Selection (CFS)
It evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them. That is the subsets of features that are highly correlated with the class while having low inter correlation are preferred. In this paper the greedy search strategy is used for WDBC data set and the random search is used for WPBC data set.

2.3Wisconsin Diagnostic and Prognostic datasets
Both the Breast Cancer Wisconsin Diagnostic dataset (WDBC) and the Breast cancer Wisconsin Prognostic (WPBC) dataset were obtained from the UCI Machine Learning Repository [9]. Features are computed from a digitized image of a Fine Needle Aspiration (FNA). The WDBC consists of 569 instances and 30 real valued input features whereas WPBC consists of 198 instances and 33 features. The attributes of the two datasets are nearly the same yet the WPBC has three additional features Time, Tumor size

and Lymph node status. The details of the attributes are given in Table1.

Table 1. Attribute information of the Breast Cancer Wisconsin datasets

Diagnostic dataset	**Prognostic dataset**
1) ID Number	ID Number
2) Diagnosis (M – Malignant, B-benign)	Outcome(R-recurrent, N- Non recurrent)
3) -----	Time (recurrence time)
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 (perimeter of the cell nucleus) d. Area (area of the cell nucleus) e. Smoothness (local variance 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 (symmetry of the cell nuclei) j. Fractal dimension (coastline approximation-1)	
34) -----	Tumor Size (size of the tumor)
35) -----	Lymph node status

3. PROPOSED METHOD

MCRSFS is a two-level feature reduction algorithm. The aim of this algorithm is to achieve a feature subset with minimum number of features providing efficient classification accuracy. This is composed of two feature reduction algorithms. Rough set QuickReduct algorithm is applied initially to obtain the minimal feature subset. Then the second algorithm CFS is used to do further reduction in minimal feature subset. Usually in QuickRreduct algorithm, the initial reduct set say R starts with an empty set but in our proposed method R is initialized with three different values as high correlation feature, average correlation feature and low correlation feature with decision attribute (class label). Because the starting point in the attribute space influences the direction of search. Then CFS is applied to the union of all the features of those three reduct sets.

The data sets are discretized for the purpose of rough set by using equal frequency with number of intervals 5 before applying the feature selection algorithms.

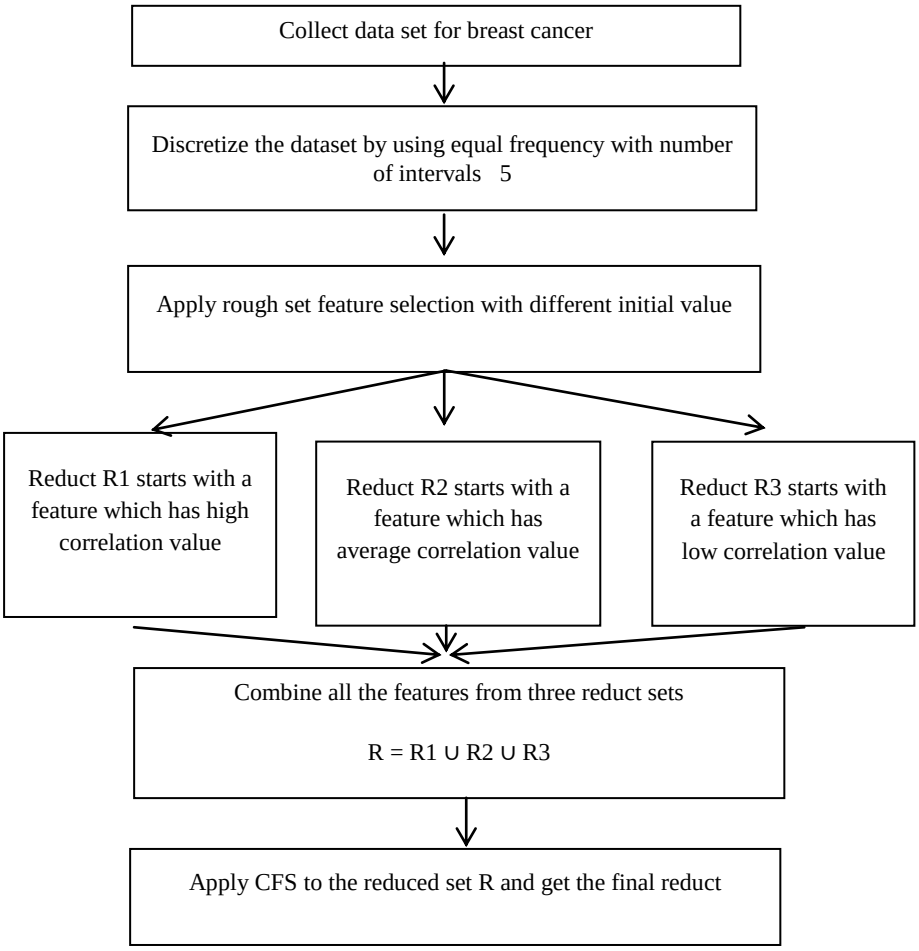

Figure 1. Proposed MCRSFS Method

Algorithm:
Input: Data set D with all features.

Output: Optimum feature subset R

Step 1: Discretize the data set using equal frequency with number of intervals 5

Step 2: Apply rough set feature selection algorithm with an initial value of the reduct set R1= {attribute which has high correlation value}.

Step 3: Apply rough set feature selection algorithm with an initial value of the reduct set R2= {attribute which has average correlation value}.

Step 4: Apply rough set feature selection algorithm with an initial value of the reduct set R3= {attribute which has low correlation value}.

Step 5: Combine all the features from R1, R2 and R3 and obtain R. i.e., R=R1∪ R2∪ R3

Step 6: CFS is applied for the reduced set R to get further reduction in the features.

4. RESULT

The proposed method has been implemented using MATLAB (Version 7.12). It is used to eliminate the unimportant and redundant features. The reduced attribute set obtained for WDBC dataset is: {11, 17, 22, 24, 25, 26, 28, and 31}. The reduced attribute set obtained for WPBC is {3, 27, and 35}. In this paper WEKA toolkit is used to analyze the datasets with the data mining algorithms [10]. Table 2 shows three different reduct sets obtained using rough set and table 3 shows the final reduct sets of WDBC and WPBC datasets after applying CFS.

Table 2. Features sequentially selected against different initial values

Data set	Reduct R1	Reduct R2	Reduct R3
WDBC	31,17,25,8,22	25,26,28,22,14	18,24,29,25,14,11
WPBC	27,9,32,15,3	28,6,3,35,31	3,19,35,16,13

Table3. Feature subsets selected on the breast cancer datasets

Data set	Combine all three reducts R1,R2 and R3	Apply CFS to get final reduct R
WDBC	8,11,14,17,18,22,24,25,26,28,29,31	11,17,22,24,25,26,28,31
WPBC	3,6,9,13,15,16,19,27,28,31,32,35	3,27,35

The data mining algorithms such as Bayes net, Naïve bayes, Multilayer perceptron, RBF network, IBK, J48 and Simple cart are used to classify WDBC and WPBC datasets with all the features and with optimum features selected by our proposed method. The results are shown in Table 4 and Table 5. To get high accuracy of a prediction model, optimal parameter setting play a crucial role. In this paper we evaluate the proper algorithmic parameters of all the mentioned eight data mining algorithms and use 80-20 training-testing partition of the data. Our results demonstrate that the proposed method improves the classification accuracy of almost all the data mining algorithms. The graphical representation of the performance of the classification algorithms of WDBC and WPBC are portrayed in Fig. 2 and Fig. 3 respectively.

Table 4. Classification Accuracy on WDBC dataset

S.No	Data Mining Algorithm	Considering all the features. (Accuracy %)	MCRSFS feature subset (Accuracy %)
1	Bayes Net	94.7368	94.7368
2	Naïve Bayes	90.3509	94.7368
3	MLP	96.4912	100
4	RBF	92.9825	99.1228
5	SMO	97.3684	97.3684
6	IBK	95.6140	98.2456
7	J48	92.9825	96.4912
8	Simple Cart	92.1053	94.7368

Table 5. Classification Accuracy on WPBC dataset

S.No	Data Mining Algorithm	Considering all the features. (Accuracy %)	MCRSFS feature subset (Accuracy %)
1	Bayes Net	77.5	82.5
2	Naïve Bayes	72.5	82.5
3	MLP	72.5	82.5
4	RBF	75	82.5
5	SMO	77.5	85
6	IBK	72.5	80
7	J48	75	82.5
8	Simple Cart	77.5	80

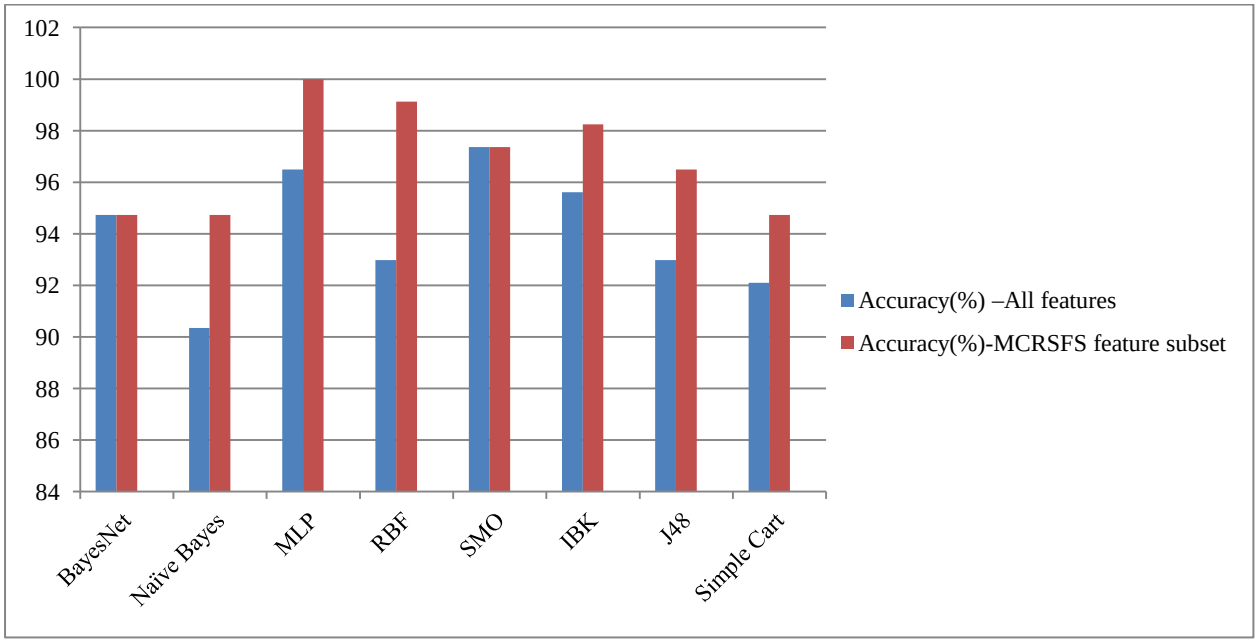

Figure 2.Classifier performance before and after feature selection on WDBC data set

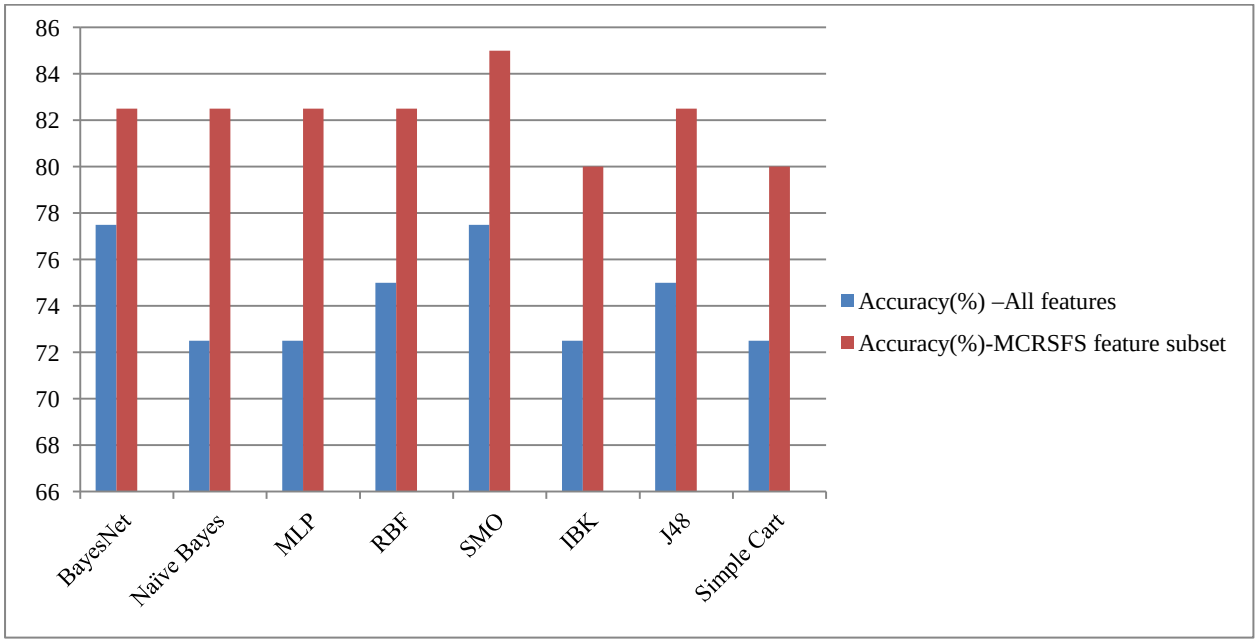

Figure3. Classifier performance before and after feature selection on WPBC data set

Comparing the accuracies in Table 4 and 5, it is found that the MCRSFS improves the accuracy of all the data mining algorithms. Multilayer perceptron has produced 100 percent accuracy in classifying the WDBC dataset and SMO is the best performing classification algorithm on the WPBC dataset which provides 85 percent classification accuracy.

Classification accuracy of MCRSFS algorithm and other methods for WDBC and WPBC from literature are summarized in Table 6.

Table 6. Accuracy rate comparison of MCRSFS with other approaches from existing researches on WDBC and WPBC datasets.

Classifier	**Data set**	**Accuracy (%)**
CART with feature selection (Chi-square) [11]	WDBC	92.61%
Hybrid Approach [12]	WDBC	95.96%
Jordan Elman neural network[13]	WDBC WPBC	98.25 70.725%
Rough set K-Means Clustering[14]	WDBC	99.12%
Proposed method(MCRSFS)	WDBC WPBC	100 85

5. CONCLUSION

In this paper MCRSFS feature selection algorithm is proposed for breast cancer datasets. This is a two level attribute reduction algorithm with the combination of rough set and CFS. The objective of this algorithm is to select minimum number of features providing high classification accuracy. It is observed that our proposed model achieved highest classification accuracy compared to other feature selection methods. MLP classification algorithm produced 100 percent accuracy in classifying the WDBC dataset. We also affirm that the SMO algorithm is the best performing algorithm on the WPBC dataset which provides 85 percent classification accuracy.

6. REFERENCES

[1] Liu H. and Motoda H., "Feature Selection for knowledge Discovery and Data Mining ", Kluwer Academic Publisher, 1999.

[2] Jensen R. and Shen Q.," A Rough Set-Aided System for Sorting WWW Bookmarks", In Zhong N *et al.* (Eds.), Web Intelligence: Research and Development, pp. 95-105, 2001.

[3] Setiono R., "Generating concise and accurate classification rules for breast cancer diagnosis", Artificial Intelligence in Medicine, 18:205–219, 2000.

[4] Chen D., Chang R.F., Huang Y.L., "Breast cancer diagnosis using self-organizing map for sonography", Ultrasound in Medical Biology 2000, Vol. 26, pp. 405–11.

[5] Giger M., Huo Z., Kupinski M., Vyborny C., "Computer-aided diagnosis in mammography. In Handbook of Medical Imaging", (Eds.) Sonka, M., Fitzpatrick, J., Medical Image Processing and Analysis, Vol. 2. SPIE Press, pp. 917–986, 2000.

[6] Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E.," A neural network approach to breast cancer diagnosis as a constraint satisfaction problem", Med. Phys. Vol.28, pp. 804–811, 2001.

[7] Zdzislaw Pawlak,"Rough Sets-Theoretical Aspects and Reasoning about Data", Klower Academic Publication. 1991..

[8] A.E.Hassanien, Z.Suraj, D.Slezak, and P.Lingras, "Rough Computing: Theories, Technologies, and Applications," NewYork: Information Science Reference, 2008.

[9]http://archive.ics.uci.edu/m1/machine-learning-databases/breast-cancer-wisconsin/breast-cancer.

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

[11] D.Lavanya, Dr.K.Usha Rani.," Analysis of feature selection with classification: Breast cancer datasets", Indian Journal of Computer Science and Engineering (IJCSE), October 2011.

[12] D. Lavanya, "*Ensemble Decision Tree Classifier for Breast Cancer Data*," International Journal of Information Technology Convergence and Services, vol. 2, no. 1, pp. 17-24, Feb. 2012.

[13] Chunekar, V.N.; Ambulgekar, H.P. (2009). "Approach of Neural Network to Diagnose Breast Cancer on Three Different Data Set," Proceedings Advances in Recent Technologies in Communication and Computing 2009 ARTcom-2009), 27th-28th Oct., IEEE, Kottayam. pp:893-895.

[14] T.Sridevii and A. Murugan, "An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set" International Journal of Computer Applications (IJCA), Vol.85, No.11, pp 38-42, Jan 2014.