An Intelligent Classifier for Breast Cancer Diagnosis based on K-Means Clustering and Rough Set

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
Feature selection aims to select subset of original features. It removes unrelated, redundant or noisy data from the problem domain. Rough set theory is often applied to feature reduction using the data alone, requiring no additional information and widely used for classification tool in data mining. Clustering, a form of data grouping, groups a set of data such that the intra-cluster similarity is maximized and the inter-cluster similarity is minimized. In this paper, k-means clustering algorithm is applied to partition the given information system and further rough set theory implemented on the data set to generate feature subset. The classification process by means of SVM is performed by using the remaining features. Wisconsin Breast Cancer datasets derived from UCI machine learning database are used for the purpose of testing the proposed hybrid model and the success rate of hybrid model is determined as 99%.

General Terms
Pattern Recognition, Machine learning.

Keywords
SVM, feature selection, rough set, clustering, classification.

1. INTRODUCTION
Feature selection (FS) is an important concept in the research of knowledge discovery**.** It aims to reduce the number of unnecessary, irrelevant or unimportant features [1]. Mining on a reduced set of features make the pattern easier to understand. The rough set theory is a process to discover redundancies and dependencies between the given features of the data to be classified. Rough set can be used in uncertainty handling and granular computing. The reduction of features is based on data dependencies [2]. The measure of dependency is calculated by a function of the approximations. Classification is a machine learning problem focusing great attention recently in the database community [3]. Rough set has been applied for classification in various applications and has been proved to be useful [4]. Clustering is the process of partitioning a set of data into a set of meaningful sub-classes that helps to understand natural grouping or structure in a data set. Objects in cluster have similar characteristics. In the field of medical diagnosis data clustering plays an important role. Breast cancer is the second fatal disease in women worldwide, early and accurate diagnosis of disease can lead to successful treatment. Classification of the nature of the disease based on the predictor feature will enable the medical experts to predict the possibility of occurrence of breast cancer in a new patient.
Our research work mainly focuses on building an efficient classifier for the Wisconsin Diagnostic Breast cancer (WDBC) and Wisconsin Prognostic Breast Cancer (WPBC) datasets from the UCI Machine learning repository. In this paper, k-means clustering is applied initially to partition the data set such that the similar characteristic objects are grouped and then by applying rough set feature selection to obtain the minimal feature subset. The classification process by means of SVM is performed by using the obtained minimal feature subset.

2. ROUGH SET BASED FEATURE SELECTION ALGORITHM
Rough set theory was introduced by Pawlak in 1982[4]. It provides many useful concepts on data mining such as representing knowledge as an equivalence class, generating rules from an information system and finding the minimal feature subset or minimal reduct. It attempts to calculate a minimal reduct without exhaustively generating all possible subsets [5]. The problem or finding minimal reduct of an information system has been the subject of much research [6]. It starts with an empty set of features. The best of the original features is determined and added to the set using the data dependency.

In rough set theory, an information table is defined as a tuple I = (U; F) where U and F are two finite, non-empty sets, U the universe of primitive objects and F the set of features. Each attribute or feature f ∈ F is associated with a set Vf of its value, called the domain of f. We may partition the feature set F into two subsets C and D, called condition and decision attributes respectively. Let P⊂F a subset of attributes. The indiscernibility relation, denoted by IND (P), is an equivalence relation defined as:

$$IND(P) = \{(x, y) \in U \times U : for\ all\ f\ \in P; f(x) = f(y)\}$$

where f(x) denotes the value of feature f of object x. If(x; y) ∈IND (P), x and y are said to be indiscernible with respect to P.

3. K-MEANS CLUSTERING
The k-means is given by MacQueen [7] and aim of this clustering algorithm is to divide the dataset into disjoint clusters. The name K-means originates from the means of the k clusters that are created from n objects.

The k-means algorithm involves randomly selecting k initial centroids where k is a user defined number of desired clusters. Each point is then assigned to a closest centroid and the collection of points close to a centroid form a cluster. The centroid gets updated according to the points in the cluster and this process continues until the points stop changing their clusters.

Given an initial set of k means (centroids) $m_1^{(1)}, \dots, m_k^{(1)}$ the algorithm proceeds by alternating between two steps:

Assignment step: Assign each observation to the cluster with the closest mean

$$S_i^{(t)} = \left\{x_j : \left\| x_j - m_i^{(t)} \right\|^2 \leq \left\| x_j - m_{i^*}^{(t)} \right\|^2 for\ all\ i^* = 1, \dots, k\right\}$$

Update step: Calculate the new means to be the centroid of the observations in the cluster.

$$m_i^{(t+1)} = \frac{1}{s_i^{(t)}} \sum_{x_j \epsilon s_i^{(t)}} x_j$$

The algorithm is deemed to have converged when the assignments no longer change.

4. PROPOSED MODEL

The proposed model is a combination of rough set feature selection method with the k-means clustering. The work of this paper is to apply the k-means clustering on an information system to partition it into two tables. Rough set feature selection can be applied on the partition tables to obtain the reduct set. After joining of reduced features from both the tables, the minimal optimal reduct is obtained. The classification process by means of SVM is performed by using the obtained minimal reduct.

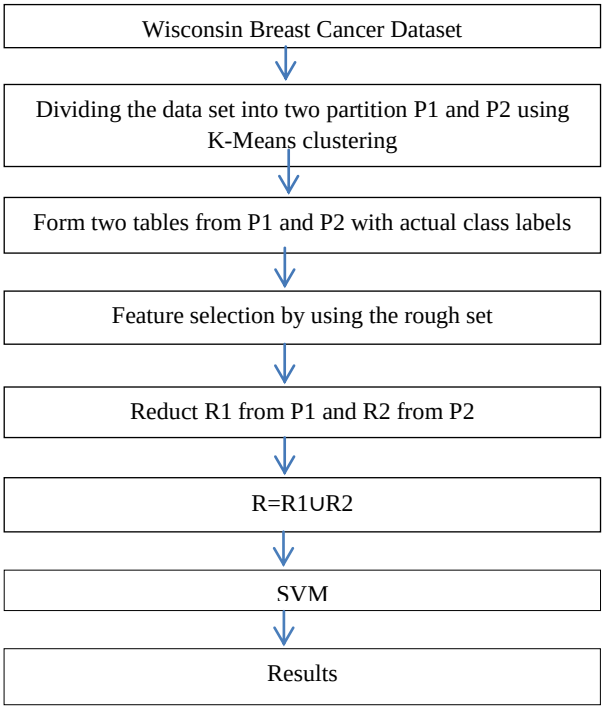

Fig 1: Proposed rough set theory and k-means clustering using SVM

5. EXPERIMENTAL ANALYSIS AND RESULT DISCUSSION

5.1 Wisconsin Diagnostic and Prognostic datasets

Both the Breast Cancer Wisconsin Diagnostic dataset (WDBC) and the Breast cancer Wisconsin Prognostic (WPBC) dataset are obtained from the UC Irvine Machine Learning Repository [8]. Features are computed from a digitized image of a Fine Needle Aspiration (FNA). The WDBC is to predict whether it is benign or malignant and WPBC is to predict whether it will recur or not. The attributes of the two datasets are nearly the same which is given in Table 1

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

5.2 Preprocessing

The missing values are replaced with appropriate values by filling the corresponding mean value. All attributes are represented in real valued measurement but they must be discretized for the purpose of rough set theory. By applying equal frequency with the number of intervals 5, the dataset is discretized and new dataset with crisp values are produced.

5.3 Feature reduction by means of rough set and k-means clustering

In the proposed method, k-mean clustering is applied to partition the given information system into two tables. Then rough set based feature selection algorithm is implemented to get the reduct R1 and R2 from the tables. Finally R1 and R2 are joined and generate the minimal feature subset.

Table 2. Sample instances of Breast cancer dataset after discretization

X	Radius	Texture	Perimeter	Area	Smoothness	Class Label
1	4	0	4	4	4	M
2	4	1	4	4	1	M
3	4	3	4	4	4	M
4	1	3	1	0	4	M
5	4	0	4	4	3	M
6	3	3	3	3	4	M
7	4	3	4	4	2	M
8	2	0	2	2	2	B
9	2	1	2	2	3	B
10	0	0	0	0	3	B

Algorithm:

Step1: Initially Table1 is partitioned into two as cluster0 and cluster1 based on k-mean clustering algorithm (sample with first 5 attributes and a decision attribute of 10 records is given in Table2). Prior to clustering the actual decision attribute is removed from the data set. On completion of the clustering process the members of each cluster are associated with their respective class label.

Step 2: Form two tables with actual class labels using two cluster groups

Step 3: Apply rough set feature selection algorithm for the dataset in Table4 and find the reduct

Step 4: Similarly find reduct set from Table5

Step5: Join the reducts generated by Table4 and Table5.

Step 6: Obtained feature subset is used as input for SVM.

Table3. Partition based on k-means clustering

X	Radius	Texture	Perimeter	Area	Smoothness	Class Label	Cluster
1	4	0	4	4	4	M	Cluster1
2	4	1	4	4	1	M	Cluster1
3	4	3	4	4	4	M	Cluster1
4	1	3	1	0	4	M	Cluster1
5	4	0	4	4	3	M	Cluster0
6	3	3	3	3	4	M	Cluster1

7	4	3	4	4	2	M	Cluster1
8	2	0	2	2	2	B	Cluster0
9	2	1	2	2	3	B	Cluster0
10	0	0	0	0	3	B	Cluster0

Table 4. Group of objects in cluster0

X	Radius	Texture	Perimeter	Area	Smoothness	Class Label	Cluster
5	4	0	4	4	3	M	Cluster0
8	2	0	2	2	2	B	Cluster0
9	2	1	2	2	3	B	Cluster0
10	0	0	0	0	3	B	Cluster0

Table5. Group of objects in cluster1

X	Radius	Texture	Perimeter	Area	Smoothness	Class Label	Cluster
1	4	0	4	4	4	M	Cluster1
2	4	1	4	4	1	M	Cluster1
3	4	3	4	4	4	M	Cluster1
4	1	3	1	0	4	M	Cluster1
6	3	3	3	3	4	M	Cluster1
7	4	3	4	4	2	M	Cluster1

Table 6.Feature subsets determined by k-means clustering and rough set

Data set	Reduct R1	Reduct R2	Reduct R
WDBC	11,5,19,21,4	26,31,5,8,15	4,5,8,11,15,19,21,26,31
WPBC	13,35,3,4	3,5,35,10	3,4,5,10,13,35

5.4 SVM Classification Results

Trials are conducted for 50-50 %, 70-30%, and 80-20% training-test partitions by using reduced feature set pertaining to the breast cancer. For all partitions ten different testing sets are conducted and the best among those were chosen. Accuracy rates obtained by means of minimal feature set are given in Table7. The WEKA tool is used to classify the data and the classification performance is evaluated by using classification accuracy and the confusion matrix. Confusion matrix at different partitions for minimal feature subset is given in Table 8. Classification accuracy of other methods for WDBC and WPBC from literature is summarized in Table 9.

Table 7.Classification accuracies for the reduced subset on different partitions.

Data set	50-50% training-test		70-30% training-test		80-20% training-test	
	All features	Reduced feature subset	All features	Reduced feature subset	All features	Reduced feature subset
WDBC	96.4789	97.8873	96.4912	98.8304	96.4912	99.1228
WPBC	73.7374	81.8182	81.3559	86.4407	80.0000	87.5

Table 8. Confusion Matrix Table At different partitions for minimal feature subset

Data set	50-50%	70-30 %	80-20%
WDBC	a b <-- classified as 94 5 | a = M 118 4 | b = B	a b <-- classified as 56 2 | a = M 0 113 | b = B	a b <-- classified as 37 1 | a = M 0 76 | b = B
WPBC	a b <-- classified as 76 3 | a = N 15 5 | b = R	a b <-- classified as 48 2 | a = N 6 3 | b = R	a b <-- classified as 32 1 | a = N 4 3 | b = R

Table 9.Accuracy rate comparison of proposed method with other approaches from existing researches on WDBC and WPBC datasets.

Classifier	Data set	Accuracy (%)
CBRGenetic [9]	WDBC	97.37
CatfishBPSO [10]	WDBC	98.17
Jordan Elman neural network[11]	WDBC WPBC	98.25 70.725%
Fuzzy rule classification [12]	WDBC	96.08
supervised fuzzy clustering [13]	WDBC	95.57%
RBF-SVM [14]	WPBC	76.32

6. CONCLUSION

This paper presents an efficient rough set feature selection based on k-means clustering algorithm for minimal reduct. The entire model has been implemented on breast cancer data sets. Using k-means clustering to generate the partition and then by applying the degree of dependency based approach of rough set theory, minimal reduct has been obtained. It is observed that the proposed model achieved the highest classification accuracies using SVM when compared with the existing methods. (99.1228 % and 87.5 % for 80-20% of training-testing partition in case of WDBC and WPBC respectively)

7. 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] Dr. DSVGK Kaladhar, Chandana B, and Bharath kumar P., "Predicting cancer survivability using classification algorithms", IJRRCS, 2(2), pp. 34-343, 2011.

[4] Ohrn A, "Rough sets: A Knowledge Discovery Technique for MultifactorMedical Outcomes", 1999.

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

[6] J.J. Alpigini, J.F. Peters, J.Skowronek, N. Shong (Eds.): "Rough sets and Current Trends in Computing", Third International Conference, RSCTC 2002. Malvern, PA, USA, October 14-16, 2002. Lecture Notes in Computer Science 2475 Springer 2002, ISBN 3-540-44274-X.

[7] MacQueen, J.B. (1967). "Some Methods for Classification and Analysis of Multivariate Observations. " In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297.

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

[9] Mohammad Darzi, Ali AsgharLiaei, Mahdi Hosseini, HabibollahAsghari. "Feature Selection for Breast Cancer Diagnosis: A Case-Based Wrapper Approach." (2011), World academy of Science, Engineering and Technology 77, 2011,pp 1142-1143.

[10] Li-Yeh Chuang, Sheng-Wei Tsai, Cheng-Hong Yang(2011), "Catfish Binary Particle Swarm Optimization for Feature Selection," Proceedings of the international Conference on Machine Learning and Computing IPCSIT vol.3 (2011)pp 40-44

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

[12] I. Gadaras, L. Mikhailov. "An interpretable fuzzy rule-based classification methodology for medical diagnosis." Artificial Intelligence in Medicine 47 (1) (2009) 25–41.

[13] J.Abonyi, and F. Szeifert. "Supervised fuzzy clustering for the identification of fuzzy classifiers." Pattern Recognition Letters, vol.14 (24), 2195–2207, 2003.

[14] Qinghua Hu, Jinfu Liu, Daren Yu."Mixed feature selection based on granulation and approximation. "Knowledge-Based System 21, 294-304.2008.