Journal of Theoretical and Applied Information Technology
31st October 2014. Vol. 68 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN INNOVATIVE ALGORITHM FOR FEATURE SELECTON BASED ON ROUGH SET WITH FUZZY C-MEANS CLUSTERING

[1] T.SRIDEVI, [2] K.SHYAMALA [2*], [3] A.MURUGAN
[1] Research Scholar, Mother Teresa Women's University, Kodaikanal, Tamil Nadu, India
[2] Associate Professor of Computer Science, Dr.Ambedkar Govt. Arts College, Chennai, Tamil Nadu, India
[3] Associate Professor of Computer Science, Dr.Ambedkar Govt. Arts College, Chennai, Tamil Nadu, India
Email:[1]sripavi_mat@yahoo.com,[3]amurugan1972@gmail.com
*Corresponding author: K.SHYAMALA, Email: shyamalakannan_2000@yahoo.com

ABSTRACT

Feature selection is a fundamental problem in data mining, especially for high level dimensional datasets. Feature selection is a process commonly used in machine learning, wherein subsets of the features from the original set of features are selected for application of a learning algorithm. The best subset contains the minimum number of dimensions retaining a suitably high accuracy on classifier in representing the original features. The objective of the proposed approach is to reduce the number of input features thus to identify the key features of breast cancer diagnosis using fuzzy c-means clustering (FCM), K-nearest neighbors (KNN) and rough set. The results show that the hybrid method is able to produce more accurate diagnosis and prognosis results than the full input model with respect to computational complexity and classification accuracy.

Keywords: *Machine Learning, Pattern Recognition, Feature Selection, FCM, Outlier Detection, Classification.*

1. INTRODUCTION

Data mining is a convenient way of knowledge extraction from high dimensional data sets. With the development of information technology, data processed by many applications is growing at an unprecedented rate, which will in turn increase the computational requirements. Data processing and knowledge discovery for massive data is always a hot topic in data mining [1]. Feature selection is an important stage of preprocessing and plays a fundamental role due to abundance of noise and irrelevant features in many real world problems. Feature selection is the process of selecting the best features among all the features because all the features are not useful in an information system and thus irrelevant or unimportant features can be eliminated without losing essential information. Traditional classification techniques do not provide better results for very high dimensional datasets. Hence feature selection is an important data preprocessing task for classification.

Rough set provides a method to decide the importance and necessity of features. It is an extension of set theory proposed by Pawlak for knowledge discovery in data sets [2]. Given a dataset with discretized features, it is possible to find a reduct of original features that are most predictive in terms of classification accuracy. Clustering is an unsupervised classification that partitions an input data set into a desired number of subgroups or clusters. It is a process of dividing the data points into their natural groups so that the data in the same group or cluster are similar to one another and different from the data points in other groups.

Outlier detection as a branch of data mining has many important applications that aim to find objects that are uncommon, deviant and exceptions [3]. Efficient detection of outliers in medical application reduces the risk of making poor decisions based on erroneous data. Detecting and eliminating such outliers may greatly enhance the performance of data mining algorithms and techniques [4].

In information retrieval, massive data and high dimensionality are two core challenges. Therefore, to find a small subset of predictive features within the data set is an appealing and encouraging tool for both challenges. In this work, FCM is used to reduce size of dataset and groups the data having similar characteristics.

Journal of Theoretical and Applied Information Technology
31st October 2014. Vol. 68 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Next deviating objects from each cluster are calculated by using the distance based function. Finally rough set feature selection is applied to find the reduct of the given data set.

The rest of the paper is organized as follows. Section 2 gives brief introduction of fuzzy c means, distance based outlier detection and rough set feature selection. The proposed method is described in section 3 and section 4 discusses the experimental results. Finally, conclusion comes in section 5.

2. METHOD DESCRIPTION

In this section, we review fuzzy *c*-means (FCM) clustering, distance-based outlier detection and rough set feature selection.

2.1 Fuzzy C Means

Clustering is an unsupervised classification mechanism where a set of data points, usually multidimensional, are classified into groups (clusters) such that members of one group are similar according to a predefined criterion [5]. FCM is the most classical method for fuzzy clustering which assigns data to multiple clusters at different degrees of membership [6, 7]. The FCM algorithm leads to minimize the following objective function:

$$J_{FCM} = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2 \quad (1)$$

where $1 \leq m < \infty$ is the fuzzifier, c_j is the j^{th} cluster center, u_{ij} is the degree of membership of x_i in the cluster j ,x_i is the i^{th} d-dimensional measured data and $\|.\|$ is the distance norm expressing the similarity between any measured data and the center.

Fuzzy partitioning is carried out through an iterative optimization of the objective function shown in equation (1), with the update of the parameters u_{ij} and the cluster center c_j by

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \quad (2)$$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} xi}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (3)$$

This iteration will stop when $\left\| u_{ij}^{(k+1)} - u_{ij}^{(k)} \right\| < \varepsilon$, where ε a termination criterion between 0 and 1, and k is the iteration step. This procedure converges to a local minimum or a saddle point of J_{FCM}.

The algorithm proceeds as follows [8]:

(i) Initialize U = u_{ij} matrix, $U^{(0)}$.

(ii) At k[th]-step calculate the center vectors $C^{(k)} = [c_j]$ with $U^{(k)}$ by (3).

(iii) Update $U^{(k)}$, $U^{(k+1)}$ by (2).

(iv) If $\left\| U^{(k+1)} - U^{(k)} \right\| < \varepsilon$ then stop; otherwise repeat steps (ii) and (iii).

2.2 Distance-Based Outlier detection

Outlier is defined as an observation that appears to be inconsistent, considerably dissimilar and exceptional with other observations in a data set. Finding anomalous points among the data points is the basic idea to find outliers and it signals out the objects mostly deviating from a given data set [9]. During clustering process irrelevant data are also allocated to some clusters that do not relate to them [10]. These outliers and noisy data should be removed in order to make more reliable clustering and data quality assurance. Isolating outliers may also have a positive impact on the results of clustering and classification and also it plays a vital role to find feature selection criterion.

For outlier detection, distance-based techniques use the distance function for relating each pair of data points of the data set. Especially top-n K[th] – nearest neighbor distance [11] is a typical top-n style outlier detection approach in distance-based techniques. For computing outlier score of object q, k numbers of nearest neighbor points of q are found first. Then average distance from object q to all k nearest neighbors is calculated. This value indicates how much an object is deviating from its neighbors. Let Nq is the set of k-nearest neighbors of object q. The k-nearest neighbor distance of q equals the average distance from q to all objects in Nq and dist(q,r) $\geq$ 0 be a

Journal of Theoretical and Applied Information Technology
31st October 2014. Vol. 68 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

distance measure between object q and r. The k-nearest neighbor distance of object q is defined as:

$$dq = \frac{1}{k}\sum_{r \in Nq} dist(q,r)$$

In this work, top-n fashion of outlier score of objects in data set is used. To reduce computational expenses, pruning is adopted before outlier detection. That is, points which are within the radius of the cluster are not considered for outlier detection because these points are very close to the centroid. So that probability of being outlier is very less for those points.

2.3 Rough set Feature Selection

In 1982, Pawlak introduced the theory of Rough sets [12, 13]. It is used to reduce original data, i.e. to find minimal sets of data with the same knowledge as in the original data and evaluates significance of data. In rough set theory, an information table is defined as I = <U,A,V,f> , U is a non-empty set of finite objects (the universe of discourse) , A is a finite set of attributes (a1,a2,….an}, which can be further divided into two disjoint subsets of C and D, A={C ∪ D } where C is condition attributes and D is a set of output or decision results. V is a set of values of attributes in C and f: C→V is the total decision function called the information function. For every set of attributes $P \subseteq C$, an indiscernibility relation IND(P) is defined in the following way: two objects x and y are indiscernible by the set of attributes P∈C if and only if $f(x, q) = f(y, q)\ \forall q \in P$.

There often exist some condition attributes that do not provide any additional information about the objects in U in the information system. So, these redundant attributes can be eliminated without losing essential classificatory information. A reduct attribute set is a minimal set of attributes from C provided that the object classification is the same as with the full set of attributes. Given C and D⊆A, a reduct is a minimal set of attributes such that

IND(C) = IND(D).

3. PROPOSED METHOD

The proposed method consists of three phases. In the first phase, a set of patterns are classified by FCM clustering. Binary classification is the task of classifying the members of a given data set into two groups on the basis of whether they have some property or not. The binary classification task in the context of medical domain is to differentiate between normal and abnormal situations. In the second phase, outliers are constructed by a distance-based technique, and finally rough set feature selection is applied to find minimal feature subset for classification. The proposed method is illustrated using Fig. 1.

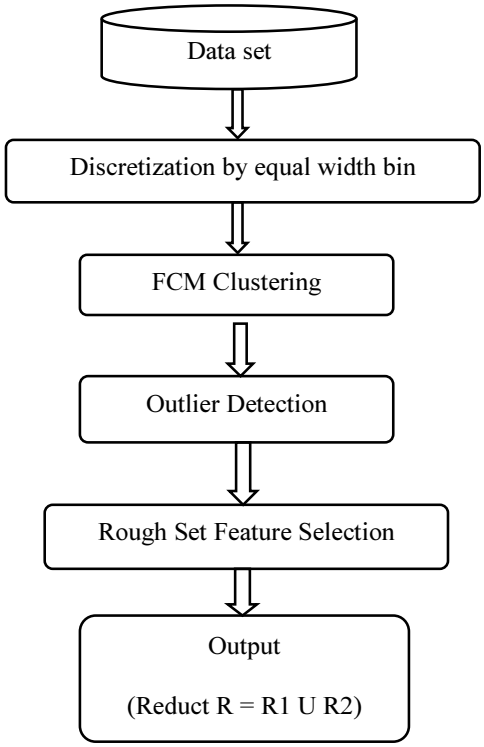

Figure 1: Block Diagram Of The Proposed Method

Algorithm:
Input : Data set D with all features.
Output : Minimal feature subset R.
Step 1: Discretize the data set using equal width binning with number of bins is 5 and fill the missing values using mean-mode method.
Step 2: Generating clusters using binary task: Initially, the entire data set is partitioned into two clusters based on FCM clustering.
Step 3: Pruning data points inside each cluster: Compute distance of each data point of a cluster from the centroid of the cluster. If the distance of the data point is less than the radius of a cluster, the point is pruned.
Step 4: Computing Outliers: Top-n K^{th} – nearest neighbor distance is adopted to compute outliers. 3% of top n points with high values are reported as outliers.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 5: Feature selection: Rough set feature selection is applied to get reduct R1 and R2 from two clusters.
Step 6: Combine all the features from R1 and R2 i.e. R = R1 U R2

4. EXPERIMENTAL RESULTS

In this experiment, Breast Cancer Wisconsin Diagnostic dataset (WDBC) and Breast Cancer Wisconsin Prognostic dataset (WPBC) are used. All the algorithms of proposed method are implemented in MATLAB 7.12.0 (R2011a). Data is collected from UCI Machine Learning Repository [14].

WDBC: This data set contains 569 medical diagnostic records, each with 32 features of attributes (ID, decisions attribute (diagnosis), and 30 real valued input features). The diagnosis is binary: Benign and Malignant.

WPBC: This data set contains 198 instances and 33 features. The attributes of this data set are nearly the same as WDBC yet it has three additional features Time, Tumor size and Lymph node status. The outcome is binary: Recurrent and Non-recurrent.

Preprocessing: The missing values are replaced with appropriate values by filling the corresponding mean-mode value. All features are represented in real valued measurement but they must be discretized for the purpose of rough set theory. By applying equal width binning with the number of bins 5, the dataset is discretized and new dataset with crisp values are produced.

Clustering: For the FCM clustering purpose only the problem part (i.e. excluding ID and outcome) of the data sets are considered. The detail of the partition is represented in Table 1.

Table 1: Number Of Data Points In Each Cluster For Cancer Data Sets

Data set	Cluster 1	Cluster 2
WDBC	316	253
WPBC	97	101

Outlier Detection: Distance based approach is applied in each cluster to find the data points those are closest to the centroid and they are pruned. Finally K-nearest neighbor is applied for remaining data points and outliers are detected based on top-n fashion distance approach. The details of outliers are summarized in table 2 and table 3.

Table 2: Number Of Data Points And 0utliers At 3% Threshold Value For WDBC Data Set

Number of data points in each cluster for WDBC		Number of outliers
Cluster 1	316	9
Cluster 2	253	7

Table 3: Number Of Data Points And Outliers At 3% Threshold Value For WPBC Data Set

Number of data points in each cluster for WPBC		Number of outliers
Cluster 1	97	3
Cluster 2	101	3

Rough set feature selection: Rough set based feature selection algorithm is implemented to get the reduct R1 and R2 from the clusters. Finally R1 and R2 are joined and the minimal feature subset is generated. The details of reduct are presented in Table 4.

Table 4: Feature subsets determined by rough set

Data set	Reduct R1	Reduct R2	Final Reduct R
WDBC	2,6,8, 15,19	5,22,24, 28,30	2,5,6,8,15,19, 22,24,28,30
WPBC	1,2,3,33	1,2,10,28,33	1,2,3,10, 28,33

To verify the effectiveness of our model, data mining algorithms such as Naïve Bayes, Multilayer perceptron, RBF network, IBK, J48 and SMO are used to classify WDBC and WPBC datasets with all the features and with minimal features selected by our proposed method. To get high accuracy of a prediction model, optimal parameter setting of classifier plays a crucial role. In this paper, we evaluate the proper algorithmic parameters of all the mentioned six data mining algorithms and use 80-20 training-testing partition of the data. The algorithmic parameters of classifiers are given in Table 5. To estimate statistical deviations the experiments are repeated for 10 runs and best among those were chosen. Our results demonstrate that the proposed method improves the classification accuracy of almost all the data mining algorithms. The comparison results of six classifiers are depicted in Table 6 and Table 7. The graphical representation of the performance of the classification algorithms of WDBC and WPBC are portrayed in Fig. 2 and Fig. 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

respectively. Time complexity before and after feature selection of both the data sets are portrayed in Fig. 4 and Fig.5.

The WEKA tool is used to classify the data and the classification performance is evaluated using classification accuracy and the time. Classification accuracy of other methods for WDBC and WPBC from literature is summarized in Table 8.

Most of the methods designed in existing algorithms use feature selection with the given training data which are available at the start of the learning process. The proposed method applies feature selection on natural grouping of data and it removes anomalous data points. Therefore, different feature subsets are generated by our method and they reduce the computational complexity of the classification algorithms.

Table 5: Parameters for Data mining algorithms

S.No.	Data mining algorithms	Parameters
1	Naïve Bayes	Default values
2	MLP	Epochs = 100, seed = 1
3	RBF	ClusteringSeed =3, numClusters=3
4	SMO	numFolds=2,randomSeed=1
5	IBK	Knn=7
6	J48	Default values

Table 6: Classification Accuracy On WDBC Data Set

S.No.	Data Mining Algorithms	Considering all the features		Feature subset obtained by the proposed method	
		Accuracy %	Time taken to build model	Accuracy %	Time taken to build model
1	Naïve Bayes	90.3509	0.02	96.3964	0.0
2	MLP	96.4912	3.84	100	0.08
3	RBF	92.9825	0.39	99.0991	0.06
4	SMO	97.3684	0.2	98.1982	0.08
5	IBK	95.6140	0.0	98.1982	0.0
6	J48	92.9825	0.14	97.2973	0.0

Table 7: Classification Accuracy On WPBC Data Set

S.No.	Data Mining Algorithms	Considering all the features		Feature subset obtained by the proposed method	
		Accuracy %	Time taken to build model	Accuracy %	Time taken to build model
1	Naïve Bayes	72.5	0.02	86.4865	0.0
2	MLP	75	1.49	89.1892	0.04
3	RBF	75	0.27	86.4865	0.01
4	SMO	77.5	0.01	86.4865	0.01
5	IBK	72.5	0.0	78.9474	0.0
6	J48	75	0.02	76.684	0.01

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

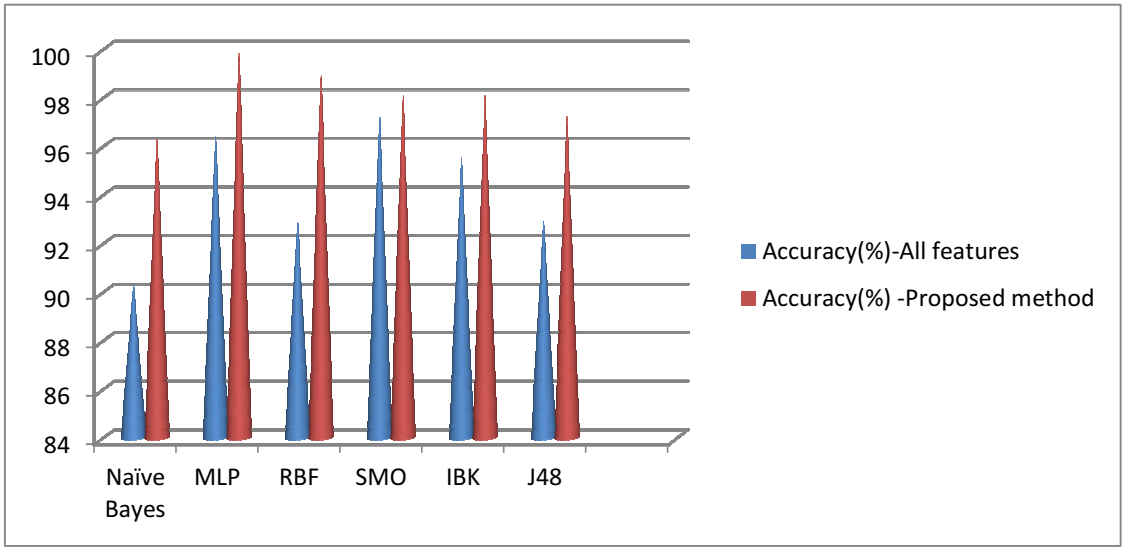

Figure 2: Classifier Performance Before And After Feature Selection On WDBC Data Set

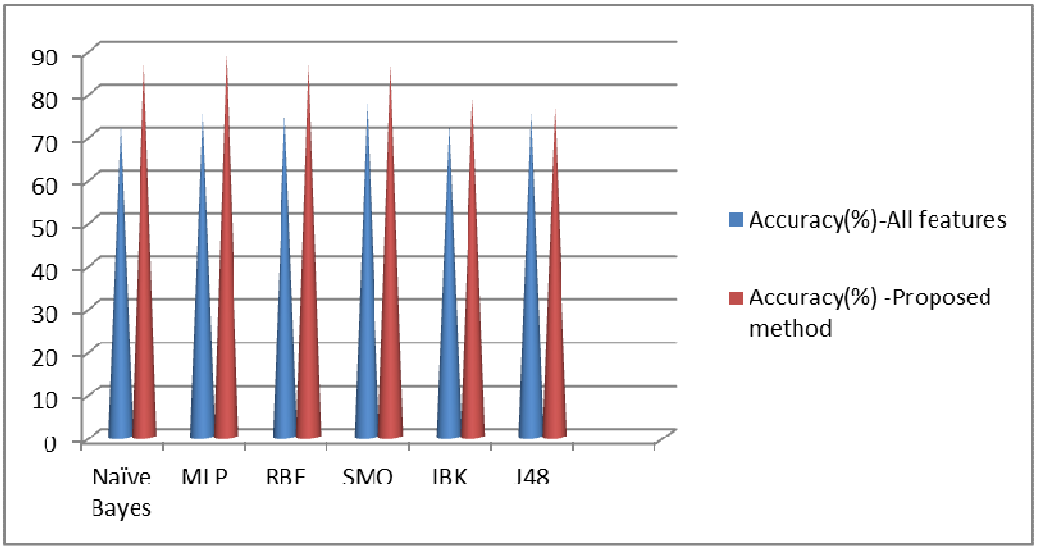

Figure 3: Classifier Performance Before And After Feature Selection On WPBC Data Set

Journal of Theoretical and Applied Information Technology
31st October 2014. Vol. 68 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

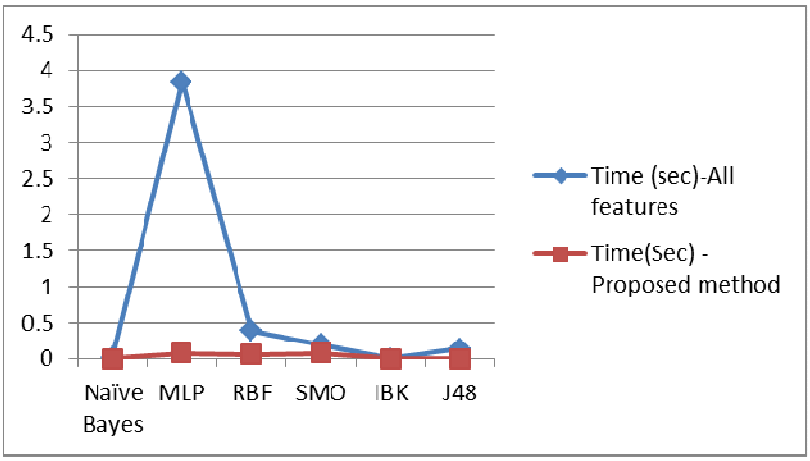

Figure 4: Time Taken To Build Model Before And After Feature Selection On WDBC Data Set

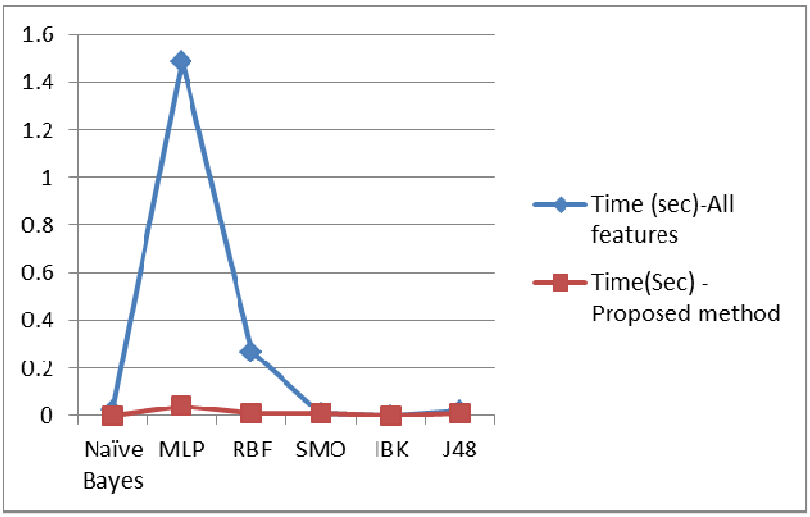

Figure 5: Time Taken To Build Model Before And After Feature Selection On WPBC Data Set

Table 8: Accuracy Rate Comparison Of Proposed Method With Other Approaches On Breast Cancer Data sets.

S.No.	Classifiers	Data set	Accuracy (%)
1.	Hybrid Approach [15]	WDBC	95.96
2.	Jordan Elman neural network[16]	WDBC WPBC	98.25 70.73
3.	RBF-SVM [17]	WPBC	76.32
4.	Rough set K-means Clustering[18]	WDBC WPDC	99.1228 87.5
5.	Modified Correlation Rough Set Feature Selection (MCRSFS) [19]	WDBC WPDC	100 85
6.	Proposed method	WDBC WPBC	100 89

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

5. CONCLUSION

This paper presents an efficient hybrid method for rough set feature selection based on FCM clustering and distanced based outlier detection. The entire model has been implemented on breast cancer data sets. Initially, FCM clustering is used to generate the partition and then by applying the distance based outlier, deviating data points have been removed. Finally, minimal feature subset has been obtained by applying degree of dependency based approach of rough set theory. Traditional feature selection algorithms find feature subset using whatever training data is given to them. The proposed method promotes the idea to actively select features from natural grouping of data and it avoids anomalous data points. Hence, the reduct obtained by our method has a positive impact on the results of classification algorithms while compared to other feature selection methods. We also affirm that the MLP algorithm is the best performing algorithm which provides 100 percent and 89 percent accuracy in classifying the WDBC and WPBC data sets respectively.

In future studies, it is possible to compare the efficacy of other classifiers by using the proposed method. Furthermore, classifiers can be applied on datasets with more features.

REFERENCES:

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, second edition, Morgan Kaufan, San Francisco, 2006.

[2] Pawlak, Z. (2002) 'Rough Sets and Intelligent Data Analysis', Information Sciences, Vol. 147, pp. 1–12.

[3] M.O. Mansur, and M. Noor Md. Sap, Outlier Detection Technique in Data Mining: A Research Perspective, In: *Postgraduate Annual Research Seminar* 2005.

[4] Hadi, Ali S., A. H. M. Imon, and Mark Werner. "Detection of outliers." *Wiley Interdisciplinary Reviews: Computational Statistics* 1.1 (2009): 57-70

[5] Pakhira, Malay K. "A modified k-means algorithm to avoid empty clusters. *"International Journal of Recent Trends in Engineering* 1.1 (2009): 220-226.

[6] Dunn, Joseph C. "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." (1973): 32-57.

[7] Bezdek, James C. "A review of probabilistic, fuzzy, and neural models for pattern recognition." *Journal of Intelligent and Fuzzy Systems* 1.1 (1993): 1-25.

[8] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," *Fuzzy Sets and Systems*, vol. 149, pp. 5-20, 2005.

[9] F. Angiulli and C. Pizzuti. Fast outlier detection in high dimensional spaces. In *PKDD '02: Proceedings of the 6th EuropeanConference on Principles of Data Mining and Knowledge Discovery*,pages 15–26, London, UK, 2002. Springer-Verlag.

[10] R. Pamula, J. K. Deka, and S. Nandi. An outlier detection method based on clustering. In *Proceedings of the 2011 SecondInternational Conference on Emerging Applications of InformationTechnology*, EAIT '11, pages 253–256, Washington, DC, USA, 2011. IEEE Computer Society.

[11] F. Angiulli and F. Fassetti, "Detecting Distance-based Outliers in Streams of Data," In Proceedings of CIKM'07, Pages 811-820, November 6-10 2007.

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

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

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

[15] 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.

[16] 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.

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

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[18] T.Sridevi 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.

[19] T. Sridevi and A. Murugan. "A Novel Feature Selection Method for Effective Breast Cancer Diagnosis and Prognosis." International Journal of Computer Applications (IJCA), Vol.88, No.11, pp 28-33, Feb 2014.