INTELLIGENT AUTOMATION & SOFT COMPUTING, 2017
http://dx.doi.org/10.1080/10798587.2017.1307626

An Intelligent Incremental Filtering Feature Selection and Clustering Algorithm for Effective Classification

U. Kanimozhi and D. Manjula

Department of Computer Science and Engineering, College of Engineering Guindy, Anna University, Chennai, Tamil Nadu, India

ABSTRACT

We are witnessing the era of big data computing where computing the resources is becoming the main bottleneck to deal with those large datasets. In the case of high-dimensional data where each view of data is of high dimensionality, feature selection is necessary for further improving the clustering and classification results. In this paper, we propose a new feature selection method, Incremental Filtering Feature Selection (IF^2S) algorithm, and a new clustering algorithm, Temporal Interval based Fuzzy Minimal Clustering (TIFMC) algorithm that employs the Fuzzy Rough Set for selecting optimal subset of features and for effective grouping of large volumes of data, respectively. An extensive experimental comparison of the proposed method and other methods are done using four different classifiers. The performance of the proposed algorithms yields promising results on the feature selection, clustering and classification accuracy in the field of biomedical data mining.

KEYWORDS

Unsupervised Classification; Fuzzy C-Means; Cluster; Fuzzy Rough sets; Feature Selection

1. Introduction

With the rapid technological evolution, large amounts of data are being continuously generated. Sources of data such as; commercial interactions, including financial transactions, search histories, product information, medical records, population databases, weather predictions, and bioinformatics data and so on, are only some examples of this landscape of data deluge. Both the size and the dimension of these data are increasing at an unprecedented rate, which has resulted in large-scale data with high dimensions. These progressively generated big datasets can be considered as valuable resources, since they can provide key insights into human behavior, market trends, diseases, engineering safety, environmental change, etc.

Feature selection for large-scale data sets has been conceived as a significant dimensional reduction technique in machine learning. It aims to improve the accuracy and performance of classifiers by removing redundant features and selecting informative features from the data. In the process of feature selection, feature evaluation criteria are used to evaluate the quality of the candidate subsets. For a feature subset, different evaluation criteria may give different results as there are five kinds of evaluation criteria such as; distance measures, information measures, dependency measures, consistency measures, and classification error rate measures. The first four evaluation criteria are used to evaluate feature subsets according to inherent characteristics of the data. The last one relies on a classification algorithm to evaluate and select useful features and is usually used to improve the classification performance, but it is time-consuming. Hence, there is a need for an efficient algorithm for selecting informative features over large-scale datasets, by decomposing large datasets and fusing it.

Clustering is one of the primary tasks used in pattern recognition and data mining communities to search large databases for various applications, and so, clustering algorithms that scale well to big data are important and useful. Cluster analysis is a method of clustering data sets with the most similarity in the same cluster and the greatest dissimilarity between different clusters. Clustering algorithms can be used to uncover unknown relations existing in a set of unlabeled data. This useful tool for data analysis has become a branch of statistical multivariate analysis and unsupervised learning for pattern recognition. In general, clustering methods can be divided into two categories; probability model-based approaches, and non-parametric approaches. Probability model-based approaches assume that the data set follows a mixture of probability distributions so that the expectation-maximization (EM) algorithm can be used as an estimation method for clustering (Isabel Timón, 2016). In non-parametric approaches, a clustering method may be based on an objective function of similarity or dissimilarity measures. As a result, partitional clustering is generally used. The most frequently used partitional methods are k-means (Hartigan and Wong, 1979) and fuzzy c-means (FCM) (Ravi, 2012). Most of the current clustering algorithms depend on iterative procedures to find local or global optimal solutions in high dimensional datasets. Many experiments with different algorithms have to be performed to find these solutions and to study the influence of different dataset features. Hence, clustering algorithms have a high intrinsic time complexity.

Classification is the most used supervised machine learning method for accurately predicting the target class for an unlabeled sample by learning from instances described by a set of attributes and a class label. As each of the many existing classification algorithms performs poorly on some data, different attempts arose to improve the original algorithms by combining them. Many different classification approaches have been proposed and used for solving real life problems, ranging from statistical methods to machine learning techniques as

CONTACT U. Kanimozhi kanimozhiu.03@gmail.com

linear classifiers (Naive Bayes classifier and logistic regression), distance estimations (k-nearest neighbours), support vector machines, rule and decision tree based methods, and the neural networks, to name a few. The nature of most classification methods is that the final classifier is combined from multiple basic classifiers, where each one of them is constructed on specific settings or subset of instances and none of them has a full set of information about the problem.

The objective of this research work is aimed at showing that the selection of more significant features from the available raw medical dataset helps the physician to arrive at an accurate diagnosis. The primary focus is on aggressive dimensionality reduction so as to end up with an increase in the prediction accuracy. The features are subjected to a double filtration process, at the end of which, only the features that increase the accuracy and form the subset with the lowest cardinality, with their corresponding rank, are obtained. The method employs an efficient strategy of ensemble feature correlation with a ranking method. The empirical results show that the proposed Incremental Filtering Feature Selection Algorithm (IF^2SA) and Fuzzy Interval based Minimal Clustering algorithm embedded classifier model achieves remarkable dimensionality reduction and clustering in the 22 medical datasets obtained from the UCI Machine Learning repository (Hettich, 1998) and Kentridge repository (Jinyan and Huiqing, 2002).

2. Literature Review

Many works have been done in this direction by various researchers in the past. Among them, (Jensen and Shen, 2004) investigated that attribute reduction is a fuzzy rough set theory based feature selection by presenting a dependency function based-reduct and designed a heuristic algorithm to search for one of the reducts. However, it has been proven to be not convergent on many real datasets and the algorithm was restructured with efficient termination criteria to achieve the convergence on all the datasets by (Bhatt and Gopal, 2005). Uniform representations of approximation spaces and their information measures were formed (Hu, 2006), introducing probability into fuzzy approximation space, a theory about fuzzy probabilistic approximation spaces.

(Hu, 2006) defined a conditional entropy based on fuzzy rough sets to characterize the dependency function-based reduct and then used the entropy to develop a feature selection algorithm. (Aboul Ella Hassanien, 2007) introduced a hybridization scheme that combines the advantages of fuzzy sets and rough sets in conjunction with statistical feature extraction technique to classify the breast cancer images. The discernibility matrix-based algorithms are usually computationally expensive, even intolerable to operate, especially for dealing with large-scale data sets with high dimensions. To overcome this difficulty, heuristic feature selection algorithms have been developed by (Tsang, 2008) using discernibility matrix to compute all the attributes reductions and pointed out that defining fuzzy lower approximation of the dependency function may result in the consequence that the membership of the fuzzy lower approximation of an attribute subset is greater than that of the original attribute set, which is a contradiction with the idea of the rough set theory.

(Chen, 2011) introduced Gaussian kernel into fuzzy-rough sets for computing fuzzy similarity relation and developed a novel method of attribute reduction with kernel tricks. (Cornelis, 2010) presented a generalization of the classical rough set framework for data-based attribute selection and reduction using fuzzy tolerance relations and introduced the concept of fuzzy decision reducts, dependent on an increasing attribute subset measure. (Zhao, 2009) invented a special case of fuzzy-rough sets (FRS) named fuzzy variable precision rough sets (FVPRSs) by combining FRS and VPRS. They employed the discernibility matrix approach to investigate the structure of attribute reductions in FVPRS and developed an algorithm to find all reductions and obtained one near-optimal attribute reduction. (Hu, 2010) incorporated a Gaussian kernel with fuzzy-rough sets and constructed a Gaussian kernel approximation based fuzzy rough set model. They introduced a Gaussian function to compute the similarities between samples and generate fuzzy information granules for each sample to approximate the decision classes.

(Hu, 2010) also introduced fuzzy entropy to measure the uncertainty in kernel approximation. (Yao, 2014) proposed a novel variable precision (θ, σ)-fuzzy rough set model based on granular (θ, σ)-fuzzy rough sets and introduced the concept of variable precision (θ, σ)-fuzzy rough sets by considering the absolute error limit. (Qian, 2015) proposed forward approximation accelerator for combining sample reduction and dimensionality reduction and the strategy enhanced the heuristic fuzzy-rough feature selection algorithms dealing with larger data sets. (Zeng, 2015) presented fuzzy rough set approaches for incremental feature selection on Hybrid Information System (HIS) to preserve information in dynamic and hybrid environment and proposed a novel hamming distance that can deal with different types of data and applied into Gaussian kernel with FRS with updating features when a new feature is added or an old one is deleted.

Clustering has been successfully applied to the analysis of datasets from several fields such as; image processing, pattern recognition, analysis of microarray data in bioinformatics, credit card behavior modeling, etc, in order to provide valuable knowledge within these fields (Agrawal, 1998; Bezdek, 1981; Hoon, Imoto, Nolan, & Miyano, 2004; Wu & Leahy, 1993; Kultur & Caglayan MU, 2015). One of the most widely used fuzzy clustering methods is the Fuzzy C-Means (FCM) algorithm (Bezdek, Ehrlich, & Full, 1984). Some parallelization efforts have been done in the literature for FCM algorithm to deal with large datasets. (I. Timón, 2016) redefined a clustering technique called Fuzzy Minimals to enhance the classification of large datasets. They revealed that there is a linear speed-up of Parallel Fuzzy Minimal (PFM) when compared to the sequential counterpart version, keeping very good classification quality. (Havens, 2012) extended FCM clustering to very large data. They compared methods that are based on sampling followed by non-iterative extension and incremental techniques that make one sequential pass through subsets of the data and kernelized versions of FCM that provide approximations based on sampling, including three proposed algorithms. Also, they presented a set of recommendations for the use of different very large FCM clustering schemes.

(Kwok, 2002) proposed an algorithm named Parallel Fuzzy C-Means (PFCM), which is designed to run on parallel computers of the Single Program Multiple Data (SPMD) model type with the Message Passing Interface (MPI) and implemented PFCM to cluster a large data set and evaluated in terms of parallelization capability and scalability. (Modenesi, 2007) presented a PFCM cluster analysis tool, which implements the calculation of clusters' centers with the degrees of membership of records to clusters, and the determination of the optimal

number of clusters for the data. Integrated cluster validation index in the optimization process, allowing the optimization of the overall parallel process. (Rahimi, 2004) proposed a PFCM algorithm for image segmentation and evaluated against sequential algorithm by dividing the computations among the processors and minimizing the need for accessing secondary storage, and enhanced the performance and efficiency of image segmentation task.

(Ravi, 2007) proposed an efficient method to cluster data points of all the images at once. The gray level histogram is used in the FCM algorithm to minimize the time for segmentation and the space required. A parallel approach is then applied to further reduce the computation time. (Soto, 2008) proposed a new approach to obtain a convex fuzzy partition, which improves the computation of membership probabilities by a new membership function, which reflects the relative position of an object with respect to each group. However, the FCM's execution time grows exponentially with the problem size as it needs prior knowledge about the number of clusters to generate, and therefore, several executions should be done to find out the optimal number of clusters.

The remaining of the paper is organized as follows: Section 3 and 4 describes the proposed methods with the suitable algorithm. Experimental results are presented in Section 5. The paper is concluded with a mention on the future scope of this work.

3. Feature Selection

In this section, we discuss the proposed feature selection algorithm called incremental filtering feature selection (IF^{2}S) algorithm based on fuzzy rough set for effective classification. This algorithm consists of three phases. In phase I, we have used the fuzzy rough set theory for selecting the suitable subsets. In phase II, the most relevant features are selected based on mutual information. Finally, the selected features are confirmed according to the feature ranking process. The following section provides the preliminary concept of fuzzy relations and fuzzy rough sets.

3.1. Fuzzy Rough Set based Subset Selection

This section deliberates about the basic concept and definition of fuzzy relations and fuzzy rough sets. The proposed feature selection algorithm uses the fuzzy rough set for effective subset selection.

3.1.1. Fuzzy Relations

The fuzzy relations demonstrate the relationship between two sets or elements of the given values. Let U is a non-empty universe and the fuzzy power set is represent as $FR(U \times U)$. In which, $U \times U$ is a power set of the given relation and$FR$is indicates the fuzzy relation on $U \times U$ if $FR \in FR(U \times U)$, where $FR(x, y)$, measures the strength of the relationship between $x \in U$and$y \in U$.

3.1.2. Fuzzy Rough Sets

According to [13], a set of lower and upper approximation operators of a fuzzy set X, which is based on fuzzy relation FR is defined, for each $x \in U$, as:

$$FRX(x) = \inf_{y \in U}^{max} \{1 - FR(x, y), X(y)\} \quad (1)$$

$$\overline{FR}X(x) = \sup_{y \in U}^{sm} \{FR(x, y), X(y)\} \quad (2)$$

To measure the degree of x certainly belonging to X and the degree of x possibly belonging to X, respectively on which the fuzzy rough set of X is defined by $FRX, \overline{FR}X$. The earlier works on fuzzy rough sets mainly focused on constructing the approximations of fuzzy sets along the line of FR and $\overline{FR}$. In this paper, we have used the existing fuzzy rough set based feature reduction (Chen et al 2011) technique for effective subset selection in the first phase of the proposed feature selection.

3.2. Mutual Information based Feature Selection

Optimal feature selection from the subset in Phase II of the proposed feature selection algorithm is based on Mutual Information (Mohamed Bennasar, 2015). In the proposed work, the necessary features are selected from the subsets, which are selected by the fuzzy rough set based subset selection. Finally, the features are placed into two subsets such asS_i and S_j based on the dependency between the different features present in the subset. The principles of information theory are discussed focusing on entropy and mutual information and explain the reasons for employing them in feature selection. The entropy of random variable measures the uncertainty of features and an average amount of information (Cover &Thomas, 2006). The entropy of a discrete random variable $X = (x_1, x_2, \ldots, x_N)$ is denoted by $E(X)$, Where x_i refers to the possible values that X can take $E(X)$.

$$E(X) = -\sum_{i=1}^{N} p(x_i) log(p(x_i)) \quad (3)$$

Where $p(x_i)$ is the probability mass function. The value of $p(x_i)$, when X is discrete, is:

$$p(x_i) = \frac{number\ of\ instants\ with\ value\ x_i}{total\ number\ of\ instants\ (N)} \quad (4)$$

Let log = 2, so $0 \leq E(X) \leq 1$. For any two discrete random variables X and $C = (c_1, c_2, \ldots c_M)$, the joint entropy is defined as:

$$JE(X, C) = -\sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, c_j) log(p(x_i, c_j)) \quad (5)$$

Where $p(x_i, c_j)$ is the joint probability mass function of the variables X and C. The conditional entropy(CE) of the variable X given C is defined as:

$$CE(C|X) = -\sum_{j=1}^{M} \sum_{i=1}^{N} p(x_i, c_j) log(p(c_j|x_i)) \quad (6)$$

The conditional entropy is the amount of uncertainty left in C when a variable X is introduced. So, the conditional entropy $CE \leq E(x, y)$ and is equal to the entropy if and only if the two variables are independent. The relation between Joint Entropy (JE) and Conditional Entropy (CE) is as follows:

$$JE(X, C) = E(X) + E(C|X) \quad (7)$$

$$CE(X, C) = E(C) + E(X|C) \quad (8)$$

The Mutual Information (MI) is the amount of information that both variables share and is defined as:

$$MI(X;C) = H(C) - H(C|X) \tag{9}$$

The Joint Mutual Information is defined as follows:

$$JMI(X;C|Y) = E(X|C) - E(X|C,Y) \tag{10}$$

$$JMI(X,Y;C) = SI(X;C|Y) + SI(Y;C) \tag{11}$$

Now, Y is a discrete variable where, $Y = (y_1, y_2, \ldots, y_N)$. The mutual information has the amount of information, which is shared by all features and also not found within feature subsets. The high joint mutual information value indicates the more relationships between two features.

3.3. Dynamic Feature Ranking

In the proposed work, we have ranked the features based on the time, the mutual information value of each feature and the joint mutual information value of a pair of features. Features' uncertainties are tackled using the joint mutual information values between the features. The mean value of *MI* is calculated for each subset of the dataset such as S_i, S_j, S_k and S_l. And, the mutual information values are considered in the range between 0 and 1. Also, the dependency of the features in a subset is represented based on time. The information gain values between any two features with more values and normalizes its values to the range [0, 1] with value 1, which indicates that knowledge of complete prediction and the value 0 indicates that X and Y are independent. Moreover, it considers a pair of features symmetrically. Entropy-based measures require nominal features; also it is possible to apply for measuring the correlations between continuous features as well when the values are discretized properly. Therefore, it is necessary to use in this work for better ranking based on their relations. At this time, correlation based feature selection (Chen et al 2012) is utilized, which uses the best-first strategy search method for calculating the merit of feature subset. However, there is a necessity to fix the stopping criteria, due to this strictly needed constrain correlation between features, which is calculated based on Symmetrical Uncertainty according to Chen et al (2012). Feature Dependency Score (*FDS*) is calculated during the time interval $< t1, t2 >$ as follows:

$$FDS = 2.0 \times \left[\frac{MI\left(S_j, \ < t1, \ t2 >\right) + MI\left(S_i, \ < t1, \ t2 >\right) - MI(S_i \ < t1, \ t2 >, \ S_j < t1, \ t2 >)}{MI\left(S_j, \ < t1, \ t2 >\right) + MI\left(S_i, \ < t1, \ t2 >\right)} \right] \tag{12}$$

Where $\mathrm{MI}\left(\mathrm{S_j}, < t1, t2 >\right)$ and $\mathrm{E}\left(\mathrm{S_i} < t1, t2 >, \mathrm{S_j} < t1, t2 >\right)$ are defined in equations (13) and (14) as follows:

$$E\left(S_j, < t1, t2 <\right) = -\sum_{fs\varepsilon FS_j} p(S_j, < t1, t2 <) \times \log_2(p(S_j, < t1, t2 <)) \tag{13}$$

Where a realistic model of a feature $\mathrm{S_j} < t1, t2 >$ is formed by evaluating the training data during the time interval $< t1, t2 >$, considering the individual's probability values of S_j during time interval $< t1, t2 >$. A new subset S_i is worked out by partitioning the existing feature subset S_j and then the relationship between subsets S_i and S_j is given by:

$$E\left(S_i, S_j\right) = -\sum_{x\epsilon X} P(S_i < t1, t2 <) \sum_{y\epsilon Y} P(S_i < t1, t2 < / S_j < t1, t2 <) \times \log_2 P(S_i < t1, t2 < / S_j < t1, t2 <) \tag{14}$$

The proposed feature selection algorithm uses the feature dependency score of the features in a subset during a particular time interval.

Algorithm 1: Incremental Filtering Feature Selection (IF2S) Algorithm Input: *Input Data*

Output: *Best Feature Subset, Optimal Feature Set, Selected Features*

Step 1: Read the input data
Step 2: Initialize $\delta = 0.45$, *Best Feature Subset* (*BFS*) $= \{\}$, *OFS* $\leftarrow \emptyset$
Step 3: Apply fuzzy rough set for feature subset selection
Phase I: Fuzzy Rough Set based Subset Selection
Step 3.1: For each feature from F
Step 3.2: Find the fuzzy relationship (*FRX*) *for every two features using equation 1 and 2.*
Step 3.3: ***If*** *FRX* $(x_i, y_i) >$ *Threshold* ***then*** *Add these features* x_i *and* y_i *into BFS* $\delta = \delta + 0.05$
Step 3.4: Repeat step 3 until $\delta = 0.95$
Step 3.5: ***If*** *BFS* $(f_i) >$ *Threshold* ***then*** *Add the feature* f_j *into OFS*
Step 3.6: Repeat step 5 until BFS reaches empty.
Step 4: Calculate the Joint Mutual Information Value for feature selection
Phase II: Joint Mutual Information based feature selection
Step 4.1: For each feature of OFS
Step 4.2: Calculate the Joint Mutual Information value using the equation 10 and 11.Step 4.3: ***If*** $(JMI(f_i, f_j) >$ *Threshold*$)$ ***then*** *Add these two features into the Feature Set (FS)*
Step 4.4: Repeat the steps 4.2 and 4.3 until OFS is empty.
Step 5: Call the Dynamic Ranking function for ranking the features ***Phase III: Dynamic Feature Ranking***
Step 5.1: For each feature of FS
Step 5.2: Calculate the Feature Dependency Score (*FDS*) *using the equation 12 for a specified time period* t_1 *and* t_2.
Step 5.3: Sort the features based on FDS in descending order.
Step 5.4: For each feature of OFS
Step 5.5: ***If*** (*FDSvalueof* $(OFS(S_i, S_j) >$ *Threshold*$)$ ***then*** *Add the feature* S_i *into SF*
Step 5.6: Repeat the step 5.4 & 5.5 until the set OFS is empty.
Step 6: Display the selected features.

4. Clustering

In this paper, we propose a new clustering algorithm called Temporal Fuzzy Interval based Minimal Clustering Algorithm based on the existing Fuzzy Minimal algorithm (Isabel Timón, 2016). This clustering algorithm uses the fuzzy rules, Euclidean distance metric for finding distance and time for effective grouping of the given data.

4.1. Temporal Fuzzy Minimal Clustering Algorithm

The Fuzzy Minimals (*FM*) algorithm proposed by Flores-Sintas et al (2001), they demonstrated that *FM* algorithm satisfies the expected characteristics of a classification algorithm in terms of scalability, adaptability, self-driven, stability and data-independent. Fuzzy clustering techniques like Fuzzy C-Means (FCM) algorithm (Bezdek, 1984) minimize an objective function that determines the prototypes of each cluster. Let *DP* be a set of n data points,

$$DP = \{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^F \tag{15}$$

Where F is the dimension of the vector space. *FM* algorithm uses the following objective function:

$$J_{(v)} = \sum\nolimits_{x\in X} \frac{d_{xv}^2}{1 + r^2 d_{xv}^2} < t_1, \ t_2 > \tag{16}$$

Where d_{xv}^2 is the Minkowski distance norm that determines the distance between two points in the dataset. The factor r measures the isotropy in the dataset. The use of Minkowski distance measure implies that we are assuming the homogeneity and isotropy of the feature space. If the homogeneity and isotropy are broken then the clusters are created in the features space. The factor r measures the disruption of the homogeneity and isotropy of the sample by a set of factors affecting the Minkowski distance in each group. In the *FM* algorithm, the objective function presented in Eq. 16 is reformulated as shown in Eq. (17)

$$J_{(v)} = \sum_{x \in X} \mu_{xv} . d_{xv}^2 < t_1, t_2 > \tag{17}$$

Where $\mu_{xv} = \frac{1}{1+r^2 . d_{xv}^2}$

Eq. (4) is the membership function that measures the degree of membership for a given element x to the cluster where v is the prototype. The *FM* algorithm is an iterative procedure that minimizes the objective function through Eq. (5), giving the prototypes that represent each cluster.

$$v = \frac{\sum_{x \in X} \mu_{xv}^2 . x}{\sum_{x \in X} \mu_{xv}^2} \tag{18}$$

It is an iterative process where two standard values are included in the computation. ε_1 establishes the error degree committed in the minimum estimation, and ε_2 shows the difference between potential minimums.

4.2. Temporal Interval based Fuzzy Minimal Clustering Algorithm

First of all, the algorithm reads the data set to be classified (X), initializes some structures to store prototypes (V) and clusters of prototypes (C). Then, it computes the factor r, using the whole dataset (X). Next, it divides the dataset equally among the clusters so that each cluster handles n/c data points (being n the total number of features and c the number of clusters). Once the different clusters receive the information, they proceed with the Temporal Interval based Fuzzy Minimal clustering (TIFMC) algorithm over the n/c data assigned to it and also with the r factor previously calculated by the proposed algorithm. Finally, all prototypes are gathered together into a unique group (V) before it proceeds with a k-means clustering to determine the final clustering result.

Algorithm 2: Temporal Interval based Fuzzy Minimal Clustering (TIFMC)
AlgorithmInput: *Input Dataset*

Output: *Suitable prototypes for clustering process V*

Step 1: Choose ε_1 and ε_2 standard parameters.
Step 2: Initialize $V = \{0\} \subset \mathbb{R}^F$, $T = 0$ // T is the time. F is the dimension of the vector space.
Step 3: Estimate factor r during the time interval $< t_1, t_2 >$
*Step 4: **for** $k = 1; k < n; k = k + 1$ **do** // n is the size of the dataset*
Step 5: Initialize $v_{(0)} = x_k, t = 0, E_{(0)} = 1, t_1 = 0, t_2 = 0$.
*Step 6: **while** $E_{(t_1,t_2)} \geq \varepsilon_1$ Begin*
Step 7: Time interval between t_1 and t_2 is evaluated and incremented by 1.
Step 8: Find the distance between two points using $\mu_{xv} = \frac{1}{1+r^2 d_{xv}^2}$, using $v_{(t-1)}$
Step 9: Find the prototype for the particular class in specific time using $\mu_{(t)} = \frac{\sum x \in X (\mu_{xv}^{(t)^2}) x}{(\mu_{xv}^{(t)})^2}$
Step 10: Find the prototype of each cluster for the class in specific time interval using $E_{(t)} = \sum_{\alpha=1}^{F} (v_{(t)}^{\alpha} - v_{(t-1)}^{\alpha})$ End
*Step 11: **If** sum of the prototypes difference between two groups >the potential minimal value **then**.*
Step 12: Add the particular prototype into the set V.
*Step 13: **End if***
*Step 14: **End for***
Step 15: Display the selected prototype for a class.
Step 16: Apply k-means clustering algorithm (Hartigan and Wong, 1979) with the selected prototype as input.
Step 17: Display the dataset with an optimal set of selected features.

Table 1. Dataset Description for Multiclass Data Sets: The Number of Samples, Features, and Classes, Respectively.

Dataset	No. of samples	No. of features	No. of classes
Cardiac Arrhythmia	60	7129	2
Dermatology	77	7129	2
Hepatitis	85	22,283	2
Pima Diabetes	72	7129	2
Back Ache	308	15,009	26
Biomed	174	12,533	11
Breast Cancer	97	24,481	Unknown
E-Coli	144	16,063	14
Haberman's Survival	79	2467	Unknown
Hypo-Thyroid	90	5920	5
Liver Disorder	190	16,063	14
Lung Cancer	410	12,533	2
Lymph nodes	45	4026	2
Post-operative patient	122	2619	2
Sick	111	11,340	3
Statlog Heart	34	7129	2
CNS	60	7129	2
Leukemia	110	22,278	Unknown
Leukemia-3C	72	5327	3
Leukemia-4C	72	11,225	3
SRBCT	83	2309	4
MLL	72	8359	5

5. Results and Discussion

The proposed approach has been evaluated by experiments on 22 biomedical datasets from the UCI machine learning repository (Hettich et al., 1998) and Kentridge repository (Jinyan andHuiqing, 2002).

5.1. Evaluation Metrics

This section describes in detail about the evaluation metrics, which are used in this work for measuring the performance of the proposed system. Classification accuracy is one of the most popular metrics in the classifier evaluation. It is the proportion of the number of true positives and true negatives obtained by the classification algorithm in the total number of instances, as given by Eq. (19)

$$Accuracy = \left[\frac{TP + TN}{TP + TN + FP + FN}\right] \tag{19}$$

Where, $TN, TP, FP, and FN$represent the number of true negatives, true positives, false positives and false negatives, respectively. Also, Clustering accuracy has been used as cluster validation metric to judge the quality of the cluster formation algorithm. Clustering Accuracy is defined as follows:

$$Clustering\ Accuracy = \frac{Number\ of\ Correct\ Count}{Total\ number\ of\ instance/sample} * 100 \tag{20}$$

5.2. Experimental Results

We conducted several experiments to demonstrate the effectiveness of the proposed algorithms. This Section analyzes the Incremental Filtering Feature Selection (IF^{2}S) Algorithm and Temporal Interval based Fuzzy Minimals Clustering (TIFMC) algorithm on 22 biomedical datasets. We focus on the

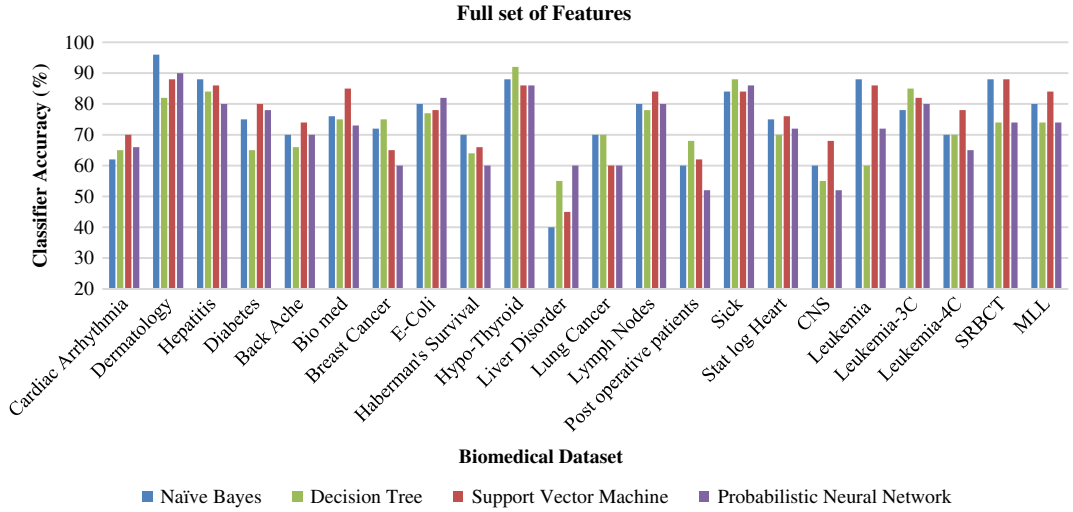

Figure 1. Performance comparison of different classifiers on a full set of features in biomedical data sets.

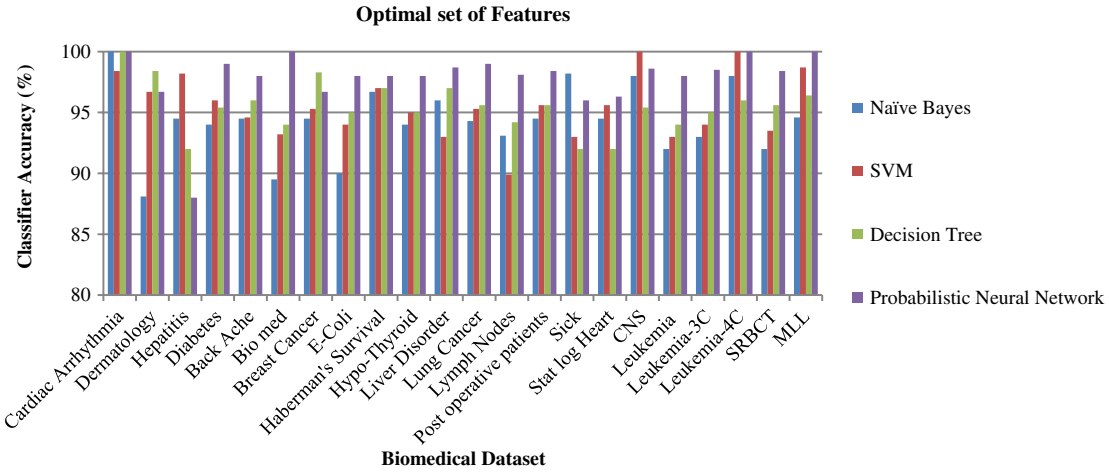

Figure 2. Performance comparisons of different classifiers on an optimal set of features selected by the proposed IF2SA in biomedical data sets.

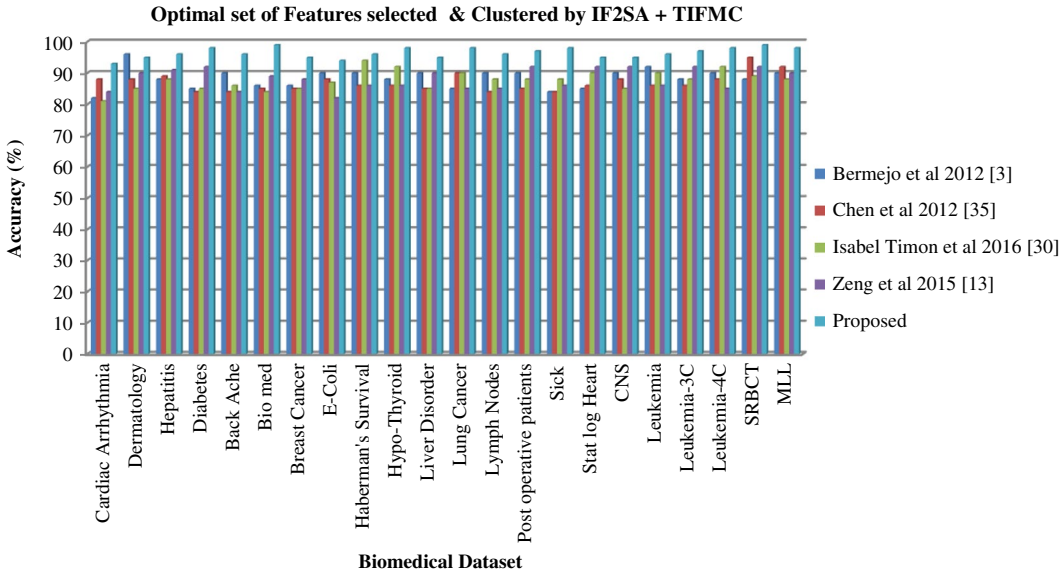

Figure 3. Classification accuracy obtained for the existing systems & the proposed IF2SA+ TIFMC by pnn classifier.

computational features of the Temporal Interval based Fuzzy Minimals Clustering algorithm and how it can be designed for handling large datasets. To guarantee the correctness of our algorithms, a quality comparison between the results obtained by other existing algorithms is also provided. We focus on performance evaluation of TIFMC algorithm using two different

Table 2. Average Values Obtained for Evaluation Metrics.

Evaluation Metrics	Value
Classification Accuracy	98.68
Sensitivity	1.0000
Specificity	0.9954
F-Measure	0.9980
Area Under Curve	1.0000

benchmarks previously described. The experiments are developed on a Windows-based laptop machine with 4 GB DDR3 memory and Intel Core I5 1.6 MHz processor using Matlab 6.0 release 12.

5.2.1. Dataset Description

We selected the UCI datasets shown in Table 1 and defined increasingly larger explorations for each dataset. Table 1 displays the datasets used in this research work. There the number of samples, the number of features, and the class distribution. When data is not available, it is represented as "unknown". In turn, Table 1 visualizes the binary and multiclass data sets. In the case of, the number of classes refers the class distribution in which a number of classes are not shown due to its high diversity.

The improvement in prediction and classification accuracy is because, the proposed TIFMC, has a minimal rough set, which is obtained by applying the clustering algorithm. Thus the experimental results show that the proposed system provides better classification accuracy. Figure 1 shows the performance comparison of different classifiers, which are used in our work to test the performance of the proposed algorithms on the full set of features available in biomedical data sets. Figure 2 shows the best average classification accuracy of the four classifiers on each dataset with the optimal set of features selected when applied to the proposed Incremental Filtering Feature Selection (IF^2S) Algorithm and the best accuracy is obtained by the Probabilistic Neural Network (PNN) classifier. Figure 3 shows the improved performance of our proposed algorithms in terms of classification accuracy. Also, provides the comparison between the existing algorithms discussed in the literature by PNN classifier. Table 2 shows the average values attained for the given performance measures.

As we have the label information of all 22 benchmark datasets, the clustering results were evaluated by comparing the obtained label of each data points with the ground truth. We used two standard measurements: The Cluster Accuracy (CA) and the normalized mutual information (NMI), higher values for both measurements will indicate good clustering performance. The visualized graph shown in Figure 4 and 5 depicts the test results of clustering accuracy and Normalized Mutual Information measure of the proposed Temporal Interval based Fuzzy Mininmals (TIFM) Clustering algorithm when compared with the existing algorithms such as; Fuzzy Minimals (FM), Parallel Fuzzy Minimals (PFM), Fuzzy C-Means (FCM) clustering algorithms, which are mentioned in our literature.

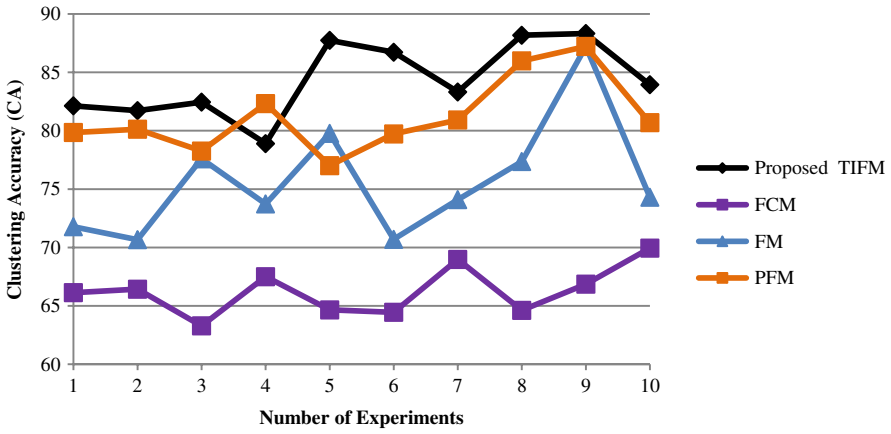

Figure 4. Comparison of clustering accuracy of the proposed tifm clustering algorithm with existing algorithms.

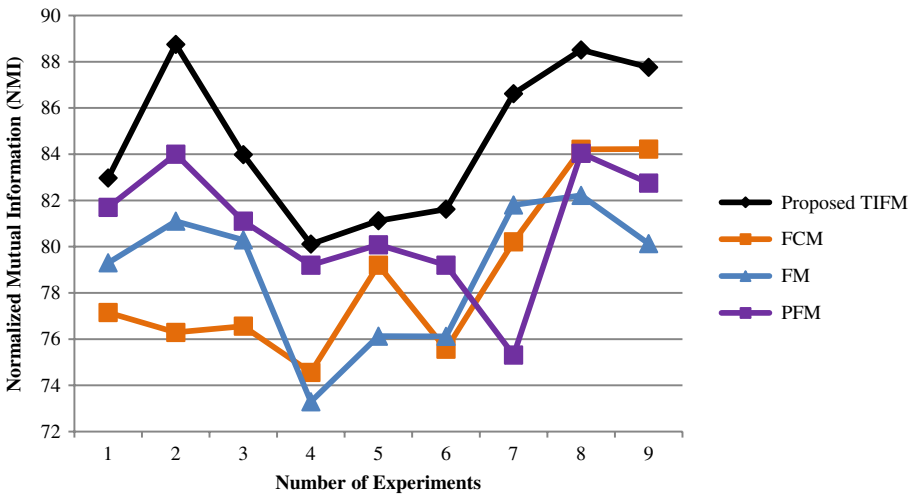

Figure 5. NMI measurement of the proposed TIFM clustering algorithm in comparison with other existing algorithms.

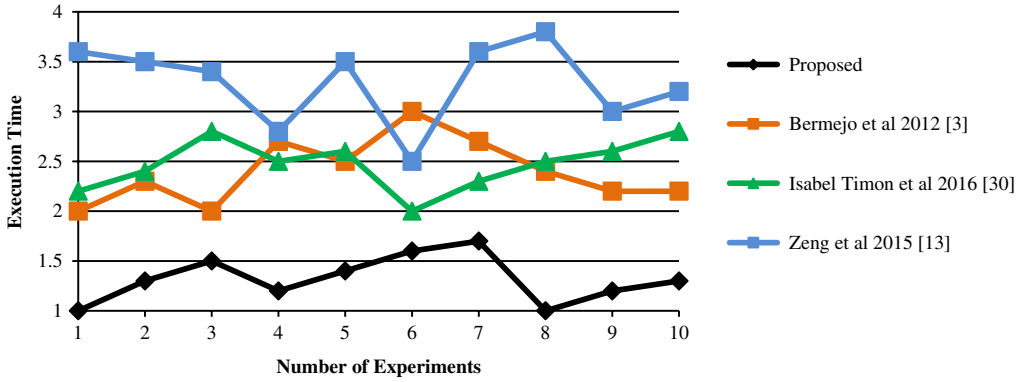

Figure 6. Execution time of the four algorithms.

5.3. Experimental Analysis of Execution Time

In this sub-section, we present the execution time to evaluate the efficiency of the proposed algorithm. The average execution time of the four algorithms is shown in Figure 6. We could derive that our algorithm is more efficient. Together, they demonstrated to be robust and efficient for exploring search spaces of possible configurations of machine learning classifiers. With the proposed methods, exploring large sets of patterns of machine learning classifiers with different data sources becomes a task that can be accomplished in reasonable and predictable time, opening the way to systematic experimentation on many of the issues around machine learning for biomedical data analysis.

At present, our work is centered on two aspects. First, dealing with the selection of an optimal set of features. Second, tuning classifier parameters is mostly a heuristic task, not existing rules providing knowledge about what parameters to choose when training a classifier. Through, we are gathering data about the performance of many classifiers, trained each one with different parameters, PNNs, SVM, etc. This by itself constitutes a dataset that can be data mined to understand what set of parameters yield better classifiers for given situations or even generally. Therefore, we intend to use the proposed framework on this bulk of classifier data to gain insight on classifier parameter tuning.

6. Conclusion and Future Work

A new feature selection method, Incremental Filtering Feature Selection (IF^{2}S) Algorithm, and a new clustering algorithm, Temporal Interval based Fuzzy Minimal Clustering (TIFMC) Algorithm, that employs the Fuzzy Rough Set for selecting an optimal subset of features and for effective grouping of a large volume of data, respectively. An extensive experimental comparison of the proposed method and other methods are done using four different classifiers. The performance of the proposed algorithms yields promising results on the feature selection, and clustering for effective improvements in classification accuracy in the field of biomedical data mining. Future works in this direction could be the introduction of intelligent agents and fuzzy rules for effective decision making.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

U. Kanimozhi, is pursuing research in the Department of Computer Science and Engineering at Anna University, Chennai. She received her Master's Degree in Computer Science and Engineering and Bachelor's Degree in Information Technology from Anna University, Chennai in 2012 and 2010 respectively. Her research interests include data mining, information retrieval, machine learning, text mining and bioinformatics, natural language processing, data analytics.

D. Manjula, is working as head and professor in the Department of Computer Science and Engineering at Anna University, Chennai. She received a Ph.D. degree in Computer Science and Engineering from Anna University of Technology, Chennai in 2004. She has published more than 200 papers in referred journals and conference proceedings. Her major research interests are big data analytics, natural language processing, cloud computing, data mining, cloud computing, social network analysis and virtualization techniques.

References

Aboul Ella Hassanien (2007). Fuzzy rough sets hybrid scheme for breast cancer detection. *Image and Vision Computing, 25*, 172–183.

Bezdek, J. C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithm*. New York, NY: Plenum Press.

Bhatt, R. B. & Gopal, M. (2005). On fuzzy-rough sets approach to feature selection. *Pattern Recognition. Letters, 26*, 965–975.

Bennasar, Mohamed, Hicks, Yulia, & Setchi, Rossitza (2015). Feature selection using Joint Mutual Information Maximisation. *Expert Systems with Applications, 42*, 8520–8532.

Bermejo, P., Ossa, L. D. L., Gamez, J. A., & Puerta, J. M. (2012). Fast wrapper feature subset selection in high dimensional datasets by means of filter re-ranking. *Knowledge-Based Systems, 25*(1), 35–44.

Chen, D. G., Hu, Q. H., & Yang, Y.-P. (2011). Parameterized attribute reduction with Gaussian kernel based fuzzy rough sets. *Information Sciences, 181*, 5169–5179.

Chen, Y. & Yu, S. (2012). Selection of effective features for ECG beat recognition based on nonlinear correlations. *Artificial Intelligence in Medicine, 54*(1), 43–52.

Cornelis, C., Jensen, R., Hurtado, G., & Ślęzak, D. (2010). Attribute selection with fuzzy decision reducts. *Inf. Sci., 180*, 209–224.

Cover, T. & Thomas, J. (2006). *Elements of information theory*. New York, NY: John Wiley & Sons.

Duranton M, Black-Schaffer D, De Bosschere K, & Maebe J. (2013). "The hipeac vision for advanced computing in horizon 2020", HiPEAC High-Performance Embedded Architecture and Compilation, pp. 1–12.

Fraley, C. & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*, 611–631.

Flores-Sintas, A., Cadenas, J. M., & Martin, F. (2001). Detecting homogeneous groups in clustering using the Euclidean distance. *Fuzzy Sets and Systems, 120*, 213–225.

Havens, T. C., Bezdek, J. C., Leckie, C., Hall, L. O., & Palaniswami, M. (2012). Fuzzy c-means algorithms for very large data. *IEEE Transactions on Fuzzy Systems, 20*, 1130–1146.

Hettich, S., Blake, C., Merz, C. (1998). UCI repository of machine learning databases. <http://www.ics.uci.edu/mlearn/MLRepository.html>

Hu, Q. H., Yu, D. R., & Xie, Z. X. (2006). Information-preserving hybrid data reduction based on fuzzy-rough techniques. *Pattern Recognition Letters, 27*, 414–423.

Hu, Q. H., Yu, D. R., Xie, Z. X., & Liu, J. F. (2006). Fuzzy probabilistic approximations spaces and their information measures. *IEEE Transactions on Fuzzy Systems, 14*, 191–201.

Hu, Q. H., Zhang, L., Chen, D. G., Pedrycz, W., & Yu, D. R. (2010). Gaussian kernel based fuzzy rough sets: Model, uncertainty measures, and applications. *International Journal of Approximate Reasoning, 51*, 453–471.

Jensen, R. & Shen, Q. (2004). Fuzzy-rough attribute reduction with application to web categorization. *Fuzzy Sets and Systems, 14*, 469–485.

Jensen, R. & Shen, Q. (2007). Fuzzy-rough sets assisted attribute selection. *IEEE Transactions on Fuzzy Systems, 15*, 73–89.

Jensen, R. & Shen, Q. (2009). New approaches to fuzzy-rough feature selection. *IEEE Transactions on Fuzzy Systems, 17*, 824–838.

Jinyan, L., Huiqing, L. (2002). Kentridge bio-medical data set repository. <http://datam.i2r.a-star.edu.sg/datasets/krbd/>

Kaufman, L. & Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kwok T, Smith K, Lozano S, & Taniar D. (2002). "Parallel fuzzy c-means clustering for large data sets", Euro-Par 2002 parallel processing, Berlin Heidelberg: Springer, pp. 365–374.

Lam, Y. K. & Tsang, P. W. M. (2012). eXploratory k-means: a new simple and efficient algorithm for gene clustering. *Applied Soft Computing, 12*, 1149–1157.

Manyika J, Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, & Byers A H. (2011). "Big data: the next frontier for innovation, competition, and productivity", McKinsey Global Institute.

Melnykov, V. & Maitra, R. (2010). Finite mixture models and model-based clustering. *Statistics Surveys, 4*, 80–116.

Modenesi, M. V., Costa, M. C., Evsukoff, A. G., & Ebecken, N. F. (2007). *Parallel fuzzy c-means cluster analysis*. Berlin Heidelberg: Springer.

Pimentel, B. A., R M C R, DSouza (2013). "A multivariate fuzzy c-means method." *Applied Soft Computing, 13*, 1592–1607.

Qian, Y. H., Wang, Q., Cheng, H. H., Liang, J. Y., & Dang, C. Y. (2015). Fuzzy-rough feature selection accelerator. *Fuzzy Sets and Systems, 258*, 61–78.

Rahimi S, Zargham M, Thakre A, and Chhillar D. (2004). "A parallel fuzzy c-mean algorithm for image segmentation", In Proceedings of the IEEE annual meeting of the fuzzy information, Processing NAFIPS'04, pp. 234–237.

Ravi A, Suvarna A, DSouza A, Reddy, G. R. M, et al. (2012). "A parallel fuzzy c means algorithm for brain tumor segmentation on multiple MRI images", In Proceedings of international conference on advances in computing, Springer India, pp. 787–794.

Soto, J., Flores-Sintas, A., & Palarea-Albaladejo, J. (2008). Improving probabilities in a fuzzy clustering partition. *Fuzzy Sets and Systems, 159*, 406–421.

Timón, Isabel, Soto, Jesús, Pérez-Sánchez, Horacio, & Cecilia, José M. (2016). Parallel implementation of fuzzy minimals clustering algorithm. *Expert Systems with Applications, 48*(15), 35–41.

Tsang, E. C. C., Chen, D. G., Yeung, D. S., Wang, X. Z., & Lee, J. W. T. (2008). Attributes reduction using fuzzy rough sets. *IEEE Transactions on Fuzzy Systems, 16*, 1130–1141.

Yao, Y. Q., Mi, J. S., & Li, Z. J. (2014). A novel variable precision (θ, σ) -fuzzy rough set model based on fuzzy granules. *Fuzzy Sets and Systems, 236*, 58–72.

Yiğit Kültür, Mehmet Ufuk Çağlayan, "A Novel Cardholder Behavior Model for Detecting Credit Card Fraud," 9th International Conference on Application of Information and Communication Technologies, 14-16 October 2015, Rostov-on-Don, Russia, pp 148–154. (DOI: 10.1109/ICAICT.2015.7338535)

Zeng, A. P., Li, T. R., Liu, D., Zhang, J. B., & Chen, H. M. (2015). A fuzzy rough set approach for incremental feature selection on hybrid information systems. *Fuzzy Sets and Systems, 258*, 39–60.

Zhao, S. Y., Tsang, E. C. C., & Chen, D. G. (2009). The model of fuzzy variable precision rough sets. *IEEE Transactions on Fuzzy Systems, 17*, 451–467.