An Efficient Clustering-Guided Fuzzy Roughset Genetic Learning for Unsupervised Feature Selection

P.Miruthula, G.S.Nandakumar

Abstract— Clustering is the application of data mining techniques to discover patterns from the datasets. Here a Fuzzy based kernel mappings clustering (FKMC) in high dimensional data is proposed which incorporates genetic roughset based feature selection concept- the process of deriving the similarity information from the unsupervised dataset. A frequent change in similarity information makes cluster aggregation a difficult task. Process of finding the optimal feature data points that are similar to a training data is challenging task which intricate linking of raw data points to one another and elimination of anomaly information .Finally, extensive experiments are governed on both synthetic and real world datasets.
Key words— Feature selection, Clustering, Rough set, Genetic algorithm, Fuzzy, correlation, Kernel mapping.

I. INTRODUCTION

Data mining, the extraction of hidden predictive information from large databases, is a strapping new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools speculate future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The converted, prospective analyses offered by data mining move beyond the analyses of past events provided by retroactive tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to reconcile. They scrub databases for hidden patterns, finding predictive statistics that authorities may miss because it lies outside their expectations.

The number of variables or features is often very high in many domains, such as image and video understanding, and data mining . Oftentimes, these high-dimensional data have many more variables than observations. In practice, not all the features are important and discriminative, since most of them are often associated or redundant to each other and sometimes noisy . These high-dimensional features may bring some disadvantages, such as over-fitting, low efficiency and poor performance, to the traditional learning models. Therefore, it is necessary and challenging to select an optimal feature subset from high-dimensional data to remove irrelevant and redundant features, increase learning accuracy

Clustering is a bisection of data into groups of similar objects. Each variety, called cluster, consists of objects that are similar in the middle of themselves and dissimilar to objects of other groups. Representing data by fewer clusters necessarily loses certain quality details but attains simplification. It represents many data objects by few clusters, and finally, it refines data by its clusters. Data shaping puts clustering in a historical perspective rooted in mathematics, statistics, and expressed as number. From the viewpoint of machine learning, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system implies a data concept. Therefore, clustering is unsupervised learning of a hidden data concept and the fuzzy clustering is the most widely used technique for hidden data analysis.

The Fuzzy clustering binary character separation described so far may not always be a convincing representation of the formation of data. Contemplate the set of two-dimensional patterns; while it can easily detect three clusters, their character is different. The first is compact, with highly concentrated patterns. The other two exhibit completely different structures. They are far less concise, with several patterns whose allocation to a given cluster may be far less certain. Mean while it may be tempted to allocate them to two clusters with varying degrees of membership. This easy and appealing idea forms a cornerstone of fuzzy sets and collections of elements with partial membership in several categories.

II. LITERATURE SURVEY

In [1] a formerly unsupervised feature selection algorithm, named clustering-guided sparse structural learning (CGSSL), is propounded by integrating cluster analysis and sparse structural analysis into a joint framework and experimentally assessed. Non negative spectral clustering is emerged to learn more accurate cluster labels of the samples which are given as input, which guide feature selection simultaneously. In [2] The two concerns involved in developing an feature subset selection algorithm for unlabeled data: the need for discovering the number of clusters in concomitance with feature selection, and the need for normalizing the bias for feature selection criteria with respect to dimension. The feature selection problem and these issues through feature optimization using Expectation-Maximization (EM) clustering and through two disparate performance criteria for evaluating candidate feature subsets. In [3] The manifold regularization medium selects features through maximizing the classification periphery between different classes and simultaneously exploiting the structure of the probability distribution that generates to both labeled and unlabeled data. To formulate the feature choosing method into a

- P.Miruthula is currently pursuing masters degree program in computer science and engineering in Kumaraguru college of technology coimbatore, India, PH:7708378904Email:palanisamymiruthula70@gmail.com
- G.S.Nanda kumar is currently working as associate professor in kumaraguru college of technology,Coimbatore,India, E-mail:nandakumar.gs.cse@kct.ac.in

convex-concave optimization problem, where the saddle point corresponds to the optimal solution. In [4] the new approach, called Multi-Cluster Feature Selection (MCFS), for unsupervised feature selection. Specifically, choose those features such that the multi-cluster structure of the data can be best preserved. The corresponding fully functional problem can be efficiently solved since it only necessitates a sparse Eigen-problem and a L1-regularized least squares problem. In [5] The concepts and algorithms of feature selection, surveys existing feature selection algorithms namely classification and clustering, groups and compares different algorithms with a categorizing framework focused on search strategies, With the categorizing framework, efforts toward building an integrated system for intelligent feature selection process. A unifying platform is proposed as an intermediate step.

III. RESEARCH METHODOLOGY

A. CORRELATION FUZZY CLUSTERING:

The objective function of Fuzzy logic is to discover the data points as cluster centroid to the optimal membership connection for estimating the centroids, and typicality is used for improving the repugnant effect of eccentricities. The function is composed of two expressions:

- The first step is the fuzzy logic function and it uses a Euclidean distance concept
- The second step is the fuzzification weighting function proponent but the two coefficients in the objective function is used only as exhibitor of membership link and typicality.

The fuzzy aggregation designates data points to c partitions by using optimal memberships. Let $X = \{x_1, x_2, x_3 \ldots x_n\}$ denote a class of data points to be partitioned into c clusters, where x_i ($i = 1, 2, 3 \ldots n$) is the data points. The goal of objective function is to discover nonlinear correspondence within the data, kernel (root) methods use embedding linking's that connect features of data to new feature spaces. The formulated technique is Fuzzy based kernel mapping (FKM) algorithm, it is an repetition clustering methodology that reduces the objective function.

Given an dataset, $X = \{x_1 \ldots x_n\} \subset Rp$, the original *KFCNC* algorithm partitions X into c fuzzy divisions by minimizing the following objective function as,

$$J(w,U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \| x_k - v_i \|^2 \quad (1)$$

Where c is the number of clusters and selected as a prescribed value, n the number of data points, u_{ik} the membership link of x_k in class i, satisfying the $\sum_{i=1}^{c} u_{ik} = 1$ the quantity scheming clustering fuzzification, and finally V is the set of cluster centres or prototypes. Here is the snapshot of obtained result

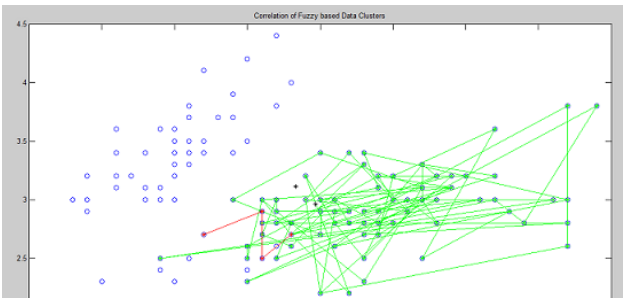

FIG: 1 CORRELATION OF FEATURES

B. ROUGH SET IN FEATURE REDUCTION

The correlation fuzzy clustering data is partitioned into three groups:

(1) A finite set of objects
(2) The group of attributes (features, variables) and
(3) The domain of attribute.

For each group in the dataset [8], a resolving system is constructed. Each decision system is subsequently split into two parts: the training and the testing dataset. Each training dataset uses the corresponding features which is given as input are made to fall into two classes: normal (+1) and abnormal (−1).

The Roughest feature equity is the process of finding a subset of features, from the aboriginal set of pattern features, optimally according to specified criterion. Rough sets theory is based on the abstract of an upper and a lower approximation of a set of elements.

An information system can be represented as,

$$S = (U, A, V, f); \quad (2)$$

where U is the universe, a finite set of N objects ($x_1, x_2, \ldots, x_N$) (a nonempty set), A is a finite set of attributes, $V = U_{a \in A} V_a$ (where V_a is a domain of the attribute a), $f : U \times A \rightarrow V$ is the net decision function (called the information function) such that $f(x, a) \in Va$ for every $a \in A$, $x \in U$. B subset of attributes $B \subseteq Q$ defines an equivalence relation (called an indiscernibility (unnoticeable) relation) on U.

$$IND(A) = \{(x, y) \in U : \text{for all } a \in B; f(x, a) = f(y, a)\}, \quad (3)$$

denoted also by A'. The system of information can also be defined as a decision table

$$DT = (U, C \cup D, V, f), \quad (4)$$

where C is a set of condition attributes, D is a set of decision attributes, $V = U_{a \in C \cup D} V_a$, where Va is the set of the domain of an attribute $a \in Q$, $f : U \times (C \cup D) \rightarrow V$ is a net decision function (knowledge function, decision rule in DT) such that $f(x, a) \in Vq$ for every $a \in A$ and $x \in V$.

The straightforward feature selection procedures are based on an evaluation of the predictive (Entropy) power of individual features, followed by a ranking of such evaluated features and eventually the choice of the first best m features. A criterion applied to an individual feature could be either of the open-loop or closed-loop type. It can be expected that a isolated feature alone may have a very low predictive power, whereas when put together with others, it may demonstrate a significant predictive power.

C. GENETIC ROUGHSET FEATURE SELECTION (GRF):

A GRF starts by generating a large set of attainable solutions to a inclined problem. It then evaluates each of those solutions, and promote on a "fitness levels" for each solution set.

The prevailing algorithm for genetic algorithms includes the succeeding steps:

Step 1: Accomplish an Initial Population. An inceptive population is fabricated from a random selection of solutions.

Step 2: Formulate fitness function. The value to fitness is assigned to each solution (chromosome) depending on how close it actually is to interpret the problem (thus arriving to the answer of the aspired problem). These "solutions" are not true value "solutions" to the problem but are possible characteristics that the methodology would employ in order to reach the answer.

Step 3: Selection, Crossover and mutation process is executed. Those chromosomes with a higher fitness value are more likely to replicate offspring, the replicated offspring is a product of the father and mother, whose make-up consists of a integration of genes from them .

Step 4: If the new inception contains a solution that assembles an output that is close enough or equal to the designated output then the problem has been solved. If it does not yield result, then the new generation will go across the same task as their parents did. This will continue till a solution is reached. Then the algorithm is over.

The obtained mean fitness values and best fitness value for specified number of iterations is displayed below

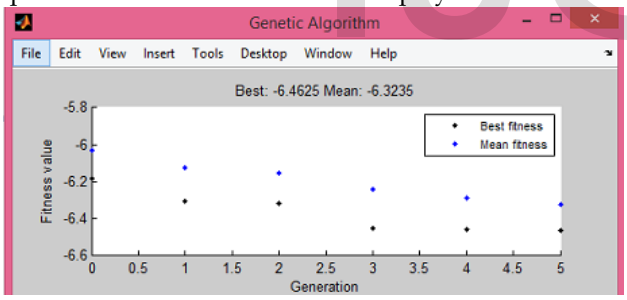

FIG: 2 GENETIC ALGORITHM

D. FUZZY BASED KERNEL MAPPING CLUSTERING ALGORITHM (FKMC):

The FKMC algorithm [9] proceeds message passing among data points. Each data points receive the availability from others data points and transmits the authority message to others data points (to pattern). Sum of responsibilities and accessibilities for data points identifies the cluster patterns. The high-dimensional data point availabilities taking A (i, k) are zero: A (i, k) = 0, R (i , k) is set to the input similarity between point i and point k as its pattern, subtracting the largest of the similitude between point i and other candidate patterns. This approach computes two kinds of notification interchanged across data points. The first one is called "responsibility" r(i, j): it is sent from data point i to the candidate paragon point j and it contemplates the accumulated evidence for how well-suited point j is to serve as the exemplar data point I. The second message is called as "availability" a(i, j) it is sent from candidate exemplar point j to point i and it reflects the accumulated evidence for how appropriate it would be for point i to choose point j as its exemplar .In the conception, the availabilities are initialized to zero: a (i, j) = 0. The update equations for r (i, j) and a (i, j) are written as

$$r(i,j) = s(i,j) - \max_{j' \neq j}\{a(i,j') + s(i,j')\} \quad (4)$$

$$a(i,j) = \begin{cases} \min\{0, r(j,j) + \sum_{i' \neq i,j} \max\{0, r(i',j)\}\}, & i \neq j \\ \sum_{i' \neq i} \max\{0, r(i',j)\}, & i = j \end{cases} \quad (5)$$

In addition, during each message's exchange between data points, a damping factor is added to avoid numerical oscillations that may arise in some circumstances:

$$R_{t+1} = (1 - \lambda)\, R_t + \lambda R_{t-1} \quad (6)$$

$$A_{t+1} = (1 - \lambda)\, A_t + \lambda\, A_{t-1} \quad (7)$$

where R = (r(i, j)) and A= (a(i, j)) represent the responsibility matrix and availability matrix, and t indicates the iteration times. The above two messages are updated iteratively, until they attain some described values or the local decisions stay constant for a number of iterations.

To calculate the distance matrix that chooses a subset of the compound space which consists only compounds which have sufficient number of close neighbors. This is obtained based on the descriptor chosen in the earlier step. The similarity measures often used in calculation of similarity between chemical compounds are Euclidean measures. The similarity measure chosen is the Euclidean distance, which is based on the triangle inequality. Euclidean measure is chosen because it shows that it was best used in shared-Neighbor clustering. Euclidean distances are usually computed from raw data and the advantage of this method is that the distance between any two object is not affected if we add new objects (such as outliers) into the analysis. The similarity measures using Euclidean distance is measured based on inter-point distance d(x1, x2) and the equations for binary descriptor is portrayed below:

$$d(x_1,x_2) = 1 - \left(\frac{\sqrt{a+b-2c}}{n}\right) \quad (8)$$

Where

a: the number of unique fragments in compound A
b: the number of unique fragments in compound B
c: the number of unique fragments shared by compounds A and B
n: the number of fragments in the compounds

The distance of the similarity matrix, the result gained will be the input for the calculation of the cluster method chosen. Finally clustered result using fuzzy concept is displayed below

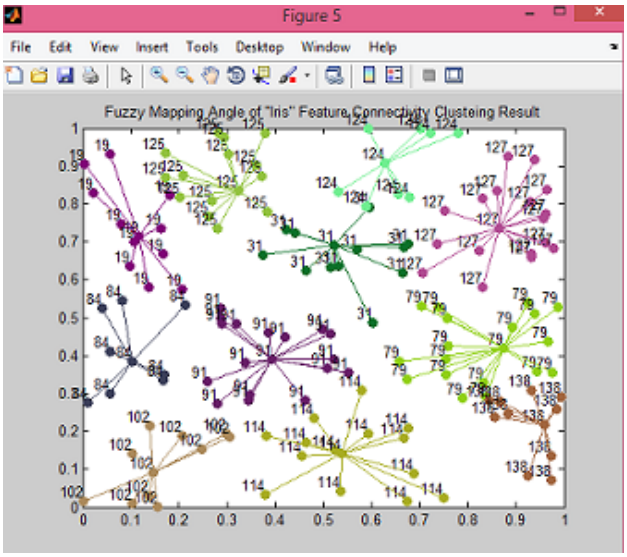

FIG: 3 FUZZY BASED CLUSTERING

III. CONCLUSION AND FUTURE WORK

Here a novel fuzzy based roughset feature selection is implemented using genetic heuristic searching algorithm for unsupervised feature selection. The proposed method can easily be extended to incorporate additional pair-wise conditions such as necessitate points with the same label to come into view in the same cluster with just an extra layer of function features. The model is flexible enough for information other than explicit constraints such as two points being in different clusters or even higher-order constraints. As a future extension PSO may be used as an optimization technique and the results can be analyzed.

REFERENCES

[1]Jennifer G. Dy and Carla E. Bradley "Feature selection for unsupervised learning" Journal of Machine Learning Research (2004)

[2] Huan Liu and Lei Yu Department of Computer Science and Engineering Arizona State University "Towards Integrating Feature Selection Algorithms for Classification and Clustering" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 17, NO. 4, APRIL 2005

[3] Zenglin Xu Rong Jin "Discriminative Semi-Supervised Feature Selection via Manifold Regularization" Journal of Machine Learning Research, 2010.

[4] Deng Cai Chiyuan Zhang Xiaofei He "Unsupervised Feature Selection for Multi-Cluster Data " International Conference on Knowledge Discovery and Data Mining 2010

[5] Zechao Li, Jing Liu, Yi Yang, Xiao fang Zhou, *Senior Member, IEEE,*and Hanqing Lu, *Senior Member, IEEE* "Clustering-Guided Sparse Structural Learning for Unsupervised Feature Selection" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 9, SEPTEMBER 2014

[6] Jiliang Tang and Huan Liu" Unsupervised feature selection framework for social media data" " IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.26,DECEMBER 2014.

[7] Haichang Li, Shiming Xiang, Zisha Zhong, Kun Ding, and Chunhong Pan " Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification" IEEE Geoscience and remote sensing letters March 2014.

[8] Roman W.WINIARSKI "Rough sets methods in feature reduction and classification" Int. J. Appl. Math. Comput. Sci., 2001, Vol.11, No.3, 565-582

[9] Richard Jensen" Combining rough and fuzzy sets for feature selection" Doctor of Philosophy School of Informatics University of Edinburgh 2005