Enhance the Performance of Clustering Technique Using Swarm Intelligence

Mr. Gajendra Dangi[1], Ms.Malti Nagle[2], Mr.Tarique Zeya Khan[3]

Research Scholar M Tech, Comp Sci. & Eng, Surabhi College Of Eng & Tech, Bhopal, M P, India[1]
Asst.Prof., M Tech, Comp Sci. & Eng, Surabhi College Of Eng & Tech, Bhopal, M P, India[2]
Asst.Prof., M Tech, Comp Sci. & Eng, Surabhi College Of Eng & Tech, Bhopal, M P, India[3]

Abstract:

In the last decade, various methods able to detect multiple clustering solutions have been introduced. According to the survey, they can bribery be categorized into methods operating on the original data-space, methods performing space transformations, and methods analysing subspace projections. The main idea is to consider each subspace as a multiple fitness constraint. For the performance evaluation of proposed algorithm used three real time datasets from UCI machine learning centre. The proposed algorithm implemented in matlab software and measures some standard parameter for the validation of proposed methodology. Our proposed method compares with two well know clustering technique such as K-means, FCM and SOC algorithm. Results shows better performance of proposed algorithm compared in existing these two algorithms

Keywords: - Clustering, SOC, PSO, FCM

I. INTRODUCTION

Clustering play an important role in discovery of unknown pattern for large database. In large database have multiple features and multiple features generate multiple views of data. In multi-view data used two clustering approach one is centralized and other is distributed approach. Centralized algorithms make use of multiple representations simultaneously to discover hidden patterns from the data. Most of the existing work in multi-view clustering follows the Centralized approach with extensions to existing clustering algorithms. Distributed algorithms first cluster each view independently from others using an appropriate single-view algorithm, and then combine the individual clustering's to produce a final partitioning. Using a partition clustering technique to generates centralized clustering process by k-means technique, but the k-means clustering technique not support multiple feature of data because it not assigned random center for cluster generation. Now in current research trend used variable weighted clustering technique for improving performance of clustering technique. in the journey of improvement of clustering technique used variable weighting clustering technique. For the more extension of clustering technique used two level weighted clustering techniques. in this dissertation proposed fuzzy based two level weighted cluster technique for multi-view data [21].

In hierarchical clustering, the data is not partitioned into clusters in a single step. Instead, a series of partitions take place, which may run from a single cluster containing all objects to clusters each containing a single object. This gives rise to a hierarchy of clustering's, also known as the cluster dendrogram.

PROPERTIES OF CLUSTERING ALGORITHMS

- Type of attributes an algorithm can handle
- Scalability to large datasets
- Ability to work with high-dimensional data
- Ability to find clusters of irregular shape
- Handling outliers
- Time complexity (we often simply use the term *complexity*)
- Data order dependency
- Labeling or assignment (hard or strict vs. soft or fuzzy)
- Reliance on a priori knowledge and user-defined parameters
- Interpretability of results

APPLICATIONS OF CLUSTERING

- Marketing: finding groups of customers with similar behavior given a large database of customer data containing their properties and past buying records.
- Biology: classification of plants and animals given their features.
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.
- City-planning: identifying groups of houses according to their house type, value and geographical location.
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds.

II. FEATURE SELECTION

A vast variety of feature selection methods have been proposed according to different metrics, such as information gain, entropy, chi-square test, t-test. Yet when applied to multi-class classification task, these methods generally suffer a pitfall of a surplus of predictive features for some classes while lack of predictive features for the remaining classes. More specifically, the strongly predictive features for the few "easy" classes rank before the weakly predictively features for the remaining "difficult" classes [13]. As a result, the features that are necessary for discriminating "difficult" classes would be ignored by traditional feature scoring methods. This problem is called the "siren pitfall". It reduces the number of features, removes irrelevant, redundant, or noisy features, and brings about palpable effects for applications: speeding up a data mining algorithm, improving learning accuracy, and leading to better model comprehensibility. Various studies show that some features can be removed without performance deterioration. Feature selection has been an active field of research for decades in data mining, and has been widely applied to many fields such as genomic analysis, text mining, image retrieval, intrusion detection, to name a few [11]. As new applications emerge in recent years, many challenges arise requiring novel theories and methods addressing high-dimensional and complex data. Feature selection for data of ultrahigh dimensionality, steam data, multi-task data, and multi-source data are among emerging research topics of pressing needs.

III. PARTICLE SWARM OPTIMIZATION

In Particle Swarm Optimization (PSO) is a swarm-based intelligence algorithm [8] influenced by the social behaviour of animals such as a flock of birds finds a food source or school of fish protecting them from a predator. A particle in PSO is analogous to a bird or fish flying through a search (problem) space. The movement of each particle is coordinated by a velocity which has both magnitude and direction. Each particle position at any instance of time is influenced by its best position and the position of the best particle in a problem space. The performance of a particles measured by a fitness value, which is problem specific. The PSO algorithm is similar to other evolutionary algorithms. In PSO, the population is the number of particles in a problem space. Particles are initialized randomly. Each particle will have a fitness value, which will be evaluated bya fitness function to be optimized in each generation. Each Particle knows its best position pbest and the best positions far among the entire group of particles gbest. The pbest of a particle is the best result (fitness value) so far reached by the particle, whereas gbest is the best particle in terms of fitness in an entire population. In each generation the velocity and the position of particles will be updated as in Eq.1 and 2, respectively. The heuristic optimizes the cost of task-resource mapping based on the solution given by particle swarm optimization technique.

$$v_i^{K+1} = \omega v_i^k + c1rand1 \times \left(pbseti - x_i^k\right) + c2rand2 \times \left(gbest - x_i^k\right) \dots\dots\dots\dots\dots (1)$$

$$x_i^{K+i} = x_i^k + v_i^{k+1} \dots\dots\dots\dots\dots\dots\dots (2)$$

Where:
v_i^k Velocity of particle i at iteration k
v_i^{K+1} Velocity of particle i at iteration k + 1
ω inertia weight
cj acceleration coefficients; j = 1, 2
randi random number between 0 and 1; i = 1, 2
x_i^k Current position of particle i at iteration k
pbesti best position of particle i
gbest position of best particle in a population
x_i^{K+i} i position of the particle i at iteration k + 1

IV. PROPOSED MODEL

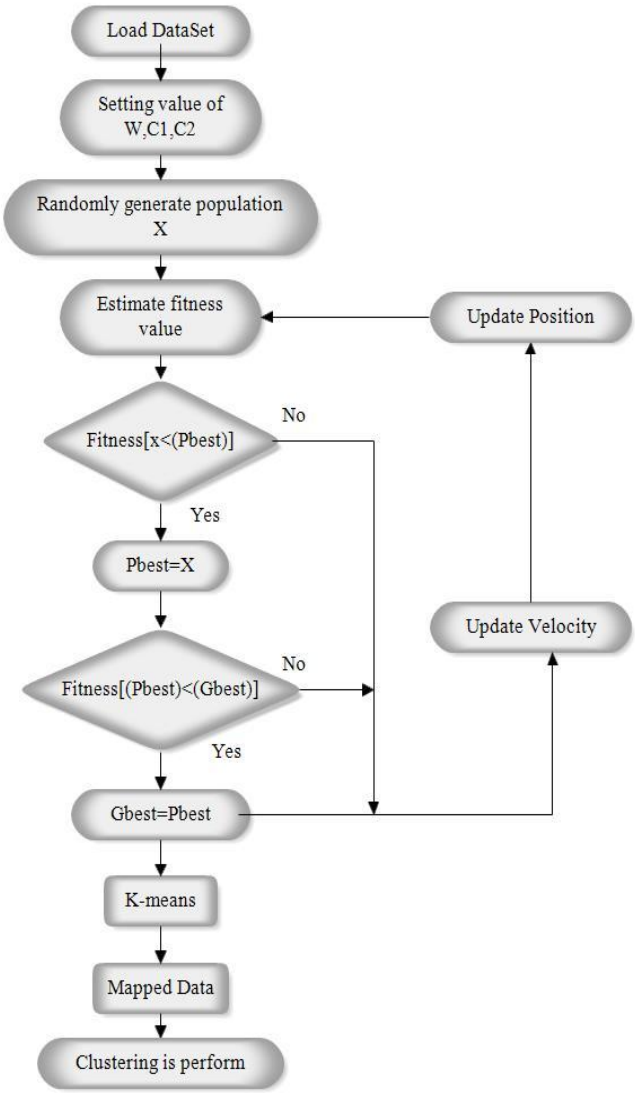

Figure 1: shows that proposed model of self-optimal clustering technique.

V. EXPERIMENTAL RESULT ANALYSIS

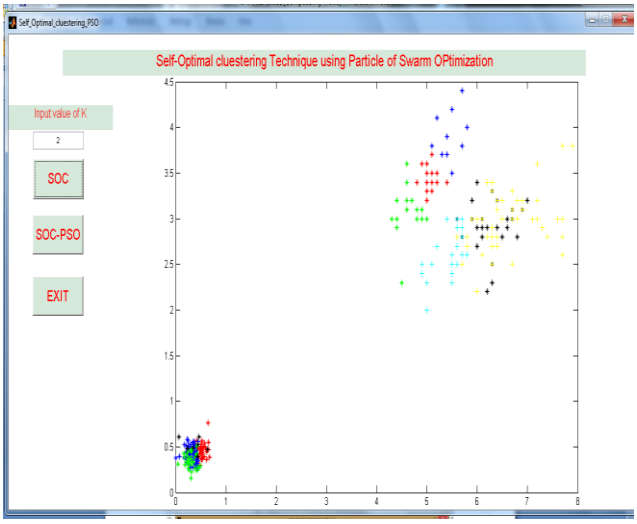

Figure 2: Shows The Output Of Input Image Ecoil Having K=2 In Soc Method.

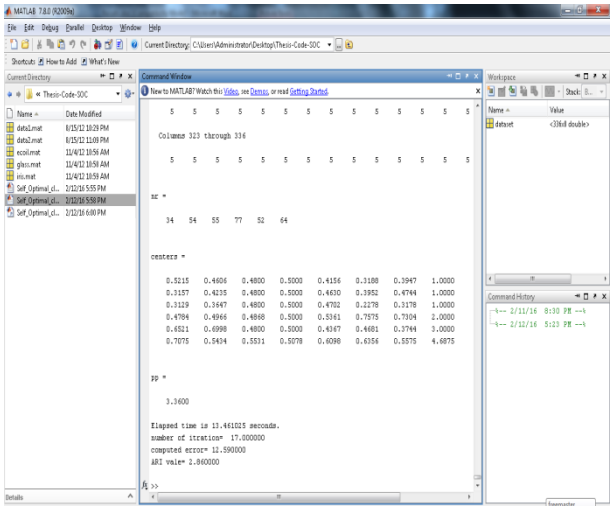

Figure 3: Shows the result of input image Ecoil having K=2 in SOC method

For the evaluation of performance of algorithm used MATLAB software and three real life data are used. The proposed algorithm work with PSO logic, so PSO function of Matlab is used. For the measuring the parameter used standard formula such as accuracy, precision, f-measure and recall.

CLUSTER-ING METHOD	GSI	PI	SI	DI	TIME
FCM	3.760	2.978	0.668	0.648	38.058
SOC	3.780	2.901	0.688	0.658	26.042
Proposed	3.800	2.387	0.748	0.718	29.144

Table 1: Shows that the performance evaluation for all clustering techniques with the input value is 2, for the Diabetes dataset.

CLUSTER-ING METHOD	GSI	PI	SI	DI	TIME
FCM	0.460	0.600	0.114	0.094	9.528
SOC	0.480	0.880	0.134	0.104	15.671
Proposed	0.500	0.722	0.194	0.164	10.293

Table 2: Shows that the performance evaluation for all clustering techniques with the input value is 4, for the Glass dataset.

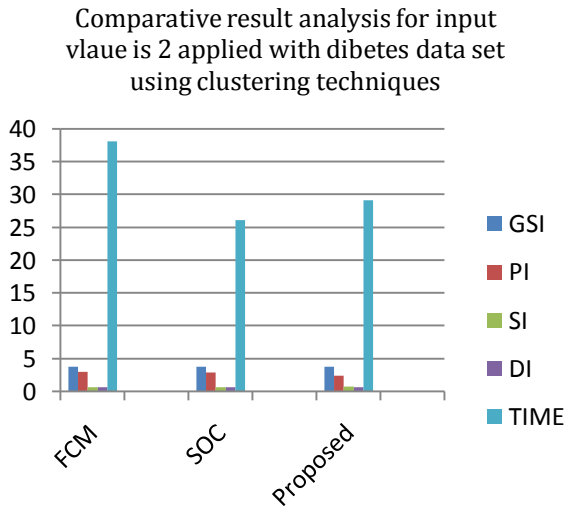

Figure 4: Shows that the comparative result for diabetes dataset using clustering techniques with the input value is 2.

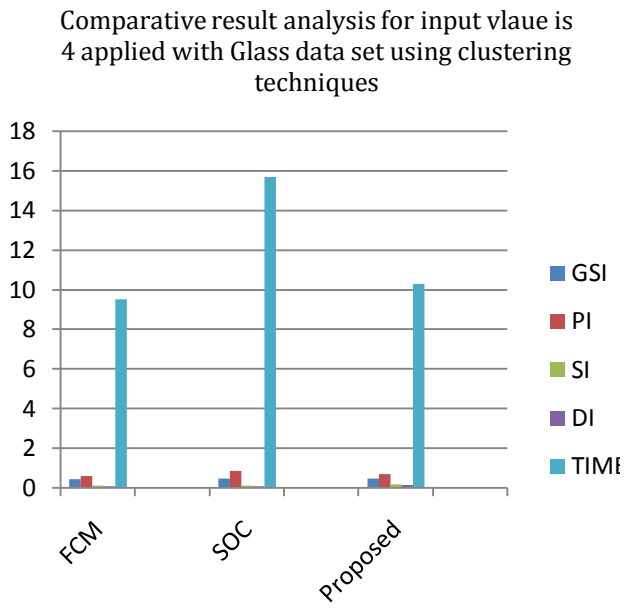

Figure 5: Shows that the comparative result for Glass dataset using clustering techniques with the input value is 4.

VI. CONCLUSION AND FUTURE WORK

In this paper, We used four datasets for the experimental process. All dataset obtained from UCI machine learning website.For the evaluation of performance of algorithm used MATLAB software and four data are used. The proposed algorithm work with PSO logic, so PSO function of MATLAB is used. For the measuring the parameter used standard formula.Our empirical result shows that our proposed algorithm shows better result in comparison of SOC-means and algorithm. The exiting two algorithms not controlled the level index of cluster and loss some data during the grouping of cluster. The proposed algorithm is very efficient for large data clustering technique.

The proposed algorithm is very efficient clustering technique for Large data. The algorithm used PSO for controlling the index variable of cluster level generation during formation of cluster. The PSO algorithm takes more time for the selection of estimated value of index. The values of index influence the cluster quality during view of data. In future reduces the computational time and complexity factor of data distribution of particle swarm optimization.

REFERENCES

[1] Nishchal K. Verma, Abhishek Roy "Self-Optimal Clustering Technique Using Optimized Threshold Function" IEEE SYSTEMS JOURNAL, IEEE 2013. Pp 1-14.

[2] Li Xuan, Chen Zhigang, Yang Fan "Exploring of clustering algorithm on class imbalanced Data" The 8th International Conference on Computer Science & Education IEEE ,2013. Pp 89-94.

[3] Ramachandra Rao Kurada, K Karteeka Pavan, AV Dattareya Rao "A preliminary survey on optimized multiobjective metaheuristic methods for data clustering using evolutionary approaches" International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, 2013. Pp 57-78.

[4] R. J. Lyon, J. M. Brooke, J. D. Knowles "A Study on Classification in Imbalanced and Partially-Labelled Data Streams" IEEE 2013. Pp 451-457.

[5] Rushi Longadge, Snehlata S. Dongre, Latesh Malik "Multi-Cluster Based Approach for skewed Data in Data Mining" IOSR Journal of Computer Engineering (IOSR-JCE) vol 12, 2013. Pp 66-73.

[6] Rukshan Batuwita, Vasile Palade "Class imbalance learning methods for support vector machines" John Wiley & Sons, Inc. 2012. Pp 1-20.

[7] M. Mostafizur Rahman and D. N. Davis "Addressing the Class Imbalance Problem in Medical Datasets" International Journal of Machine Learning and Computing, Vol. 3,2013. Pp 224-229.

[8] Nenad Tomasev,Dunja Mladeni "Hub Co-occurrence Modeling for Robust High-dimensional kNN Classification" IEEE 2009. Pp 125-141.

[9] Dech Thammasiri , Dursun Delen , Phayung Meesad , Nihat Kasap "A critical assessment of imbalanced class distribution problem: The case of predicting freshmen student attrition" Expert Systems with Applications, Elseivet ltd 2013. Pp 1220-1230.

[10] Hualong Yu, Shufang Hong, Xibei Yang" Recognition of Multiple Imbalanced Cancer Types Based on DNA Microarray Data Using Ensemble Classifiers" Hindawi Publishing Corporation BioMed Research International Volume 2013. Pp 201-214.

[11] V. Garc,J. S. S_anchez,R. Mart ,elez,R. A. Mollineda" Surrounding neighborhood-based SMOTE for learning from imbalanced data sets" Institute of New Imaging Technologies,2010. Pp 1-14.

[12] Mohammad Behdad, Luigi Barone, Mohammed Bennamoun and Tim French "Nature-Inspired Techniques in the Context of Fraud Detection" in IEEE transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 42, no. 6, november 2012.

[13] Alberto Fernandez, Maria Jose del Jesus and Francisco Herrera "On the influence of an adaptive inference system in fuzzy rule based classification system for imbalanced data-sets" in Elsevier Ltd. All rights reserved 2009.

[14] P. Garcia-Teodoro, J. Diaz-Verdejo, G. Macia-Fernandez and E.Vazquez "Anomaly-based network intrusion detection: Techniques, Systems and challenges" in Elsevier Ltd. All rights reserved 2008.

[15] Terrence P. Fries "A Fuzzy-Genetic Approach to Network Intrusion Detection" in GECCO 08, July12–16, 2008, Atlanta, Georgia, USA.

[16] Zorana Bankovic, Dusan Stepanovic,Slobodan Bojanic and Octavio Nieto-Taladriz "Improving network security using genetic algorithm approach" in Published by Elsevier Ltd 2007.

[17] Mrutyunjaya Panda and Manas Ranjan Patra "network intrusion detection using naive bayes" in

IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007.
[18] Animesh Patcha and Jung-Min Park "An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends" in Computer networks 2007.

[19] Ren Hui Gong, Mohammad Zulkernine and Purang Abolmaesumi "A Software Implementation of a Genetic Algorithm Based Approach to Network Intrusion Detection" in IEEE 2005.
[20] Jonatan Gomez and Dipankar Dasgupta "Evolving Fuzzy Classifiers for Intrusion Detection" in IEEE 2002.

BIOGRAPHY

Gajendra Dangi received the bachelor's degree in Information Technology from PATEL college of Science and Technology from RGPV University of Bhopal in 2012.He is currently Pursuing the M.Tech in computer science and engineering from SCET college From RGPV University Bhopal, M.P.

Malti Nagle received the B.E.in Information Technoloy from Samrat Ashok Technologyical Institute VIDISHA (M.P)in 2006 and M.tech in Computer Science & Engineering from Jaypee University of Information Technology(U.P)in 2009.She is a professor of Computer Science & Engg With Surabhi College Of Enginerring BHOPAL,MP.prior to that,she led the IT Trainer At Smritinet.com.BHOPAL,MP.Her main research intrests are Network security and ADHOC Network in which she has published more than 15 papers.Prof Malti was an IEEE review committee member at SoftCOM 2014.she has been in education profession from 7years.She worked in various university as A.P.