IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

# Improved the Efficiency of Self Optimal Clustering Technique using Particle Swarm Optimization

**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[2,3]

**Abstract:** In the last decade, various methods able to detect multiple clustering solutions have been introduced. According to the survey, they can briery 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 dataset from UCI machine learning center. 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, GA, SOC, PSO.

## 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]. The self-optimal clustering technique faced a problem of index generation and validation of data control. For the validation of data used swarm based optimization technique. the family of swarm intelligence gives better optimal value of index for the process of cluster generation. Hierarchical agglomerative clustering (HAC) is a bottom-up hierarchical clustering algorithm. In HAC, points are initially allocated to singleton clusters, and at each step the "closest" pair of clusters is merged, where closeness is defined according to a similarity measure between clusters. The algorithm generally terminates when the specified "convergence criterion" is reached, which in our case is when the number of current clusters becomes equal to the number of clusters desired by the user [12]. Different cluster-level similarity measures are used to determine the closeness between clusters to be merged single-link, complete-link, or group-average. Different HAC schemes have been recently shown to have well defined underlying generative models single-link HAC corresponds to the probabilistic model of a mixture of branching random walks, complete-link HAC corresponds to uniform equal-radius hyper spheres, whereas group-average HAC corresponds to equal-variance configurations. So, the HAC algorithms can be categorized as generative clustering algorithms.

## 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

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

"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 ALGORITHM

In this section discuss the proposed algorithm based on partition clustering and particle of swarm optimization. The particle of swarm optimization gives the optimal number of cluster and validate point of center and data.

Step1. Initially the data passes through the PSO and PSO define and initialized data in terms of particle and decide random size of population N=1000.

a. Define the velocity of particle in terms of data point difference value
b. Define the value of fitness constraints for the selection of data for the process of k-means algorithm

$$D_k\left(N_{i,R_j}\right) = \frac{Wij\epsilon(Cij)}{\sum_i p(i,j)}, Ri\epsilon Lk(Ci, Ri) \dots\dots\dots\dots\dots (4.5.1)$$

Here (Mi, Ri) is the value of attribute and mapping for seed

c. Iteration process is done and calculate the value of Gbest and Pbest
d. Passes data through k-means

Step2. Here show steps of processing of SOC

1) Process the PSO data and initialized the number of index.
2) Randomly select the PSO vector for the process of index optimization.
3) Every particle is examined to find the best match PSO.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

4) The similarity of index is decrease and the number of optimal PSO is going to k-means
5) After that the index value are adjusted and passes through the PSO space of cluster map.

The function PSO mapping creates data matrix of index.
1. After processing of this of PSO data creates cluster.
2. Generate PSO mapping of each cluster according to the optimal index.
3. The cluster measures the Similarity and return the equivalent cluster of data.

If the relevant cluster is not found that the process going again in PSO space.

## V. EXPERIMENTAL RESULT ANALYSIS

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.

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

| CLUSTERING METHOD | GSI | PI | SI | DI | TIME |
|---|---|---|---|---|---|
| **K-means** | 3.740 | 2.163 | 0.658 | 0.638 | 26.144 |
| **FCM** | 3.760 | 2.978 | 0.668 | 0.648 | 38.058 |
| **SOC** | 3.780 | 2.901 | 0.688 | 0.658 | 26.042 |
| **SOC-PSO** | 3.800 | 2.387 | 0.748 | 0.718 | 29.144 |

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

| CLUSTERING METHOD | GSI | PI | SI | DI | TIME |
|---|---|---|---|---|---|
| **K-means** | 0.440 | 0.647 | 0.104 | 0.084 | 10.307 |
| **FCM** | 0.460 | 0.600 | 0.114 | 0.094 | 9.528 |
| **SOC** | 0.480 | 0.880 | 0.134 | 0.104 | 15.671 |
| **SOC-MOGA** | 0.500 | 0.722 | 0.194 | 0.164 | 10.293 |

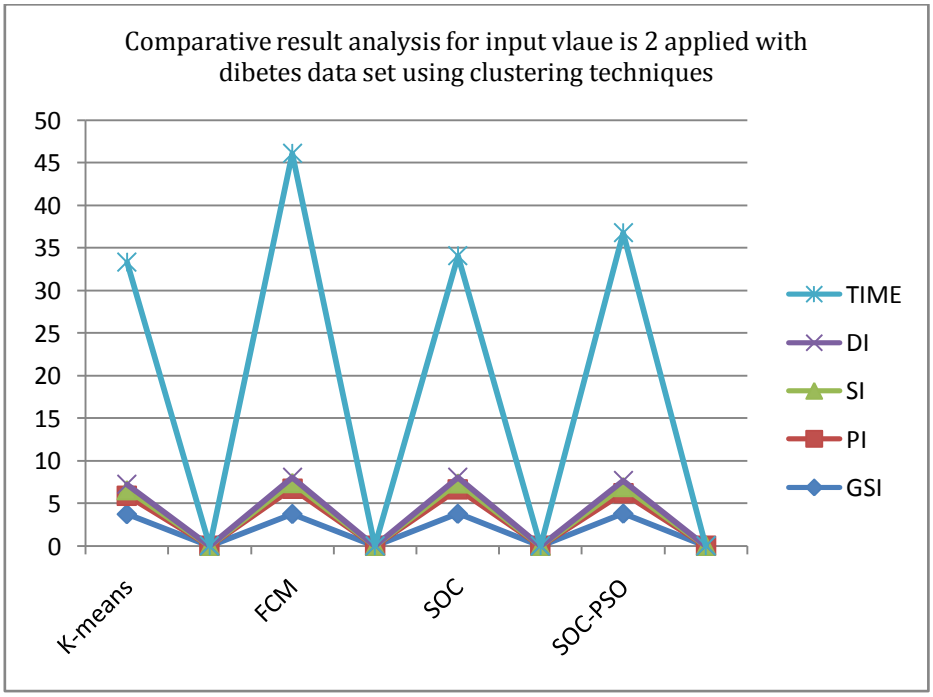


**Figure 1: Shows that the comparative result for diabetes dataset using clustering techniques with the input value is 2.**

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 7, July 2017

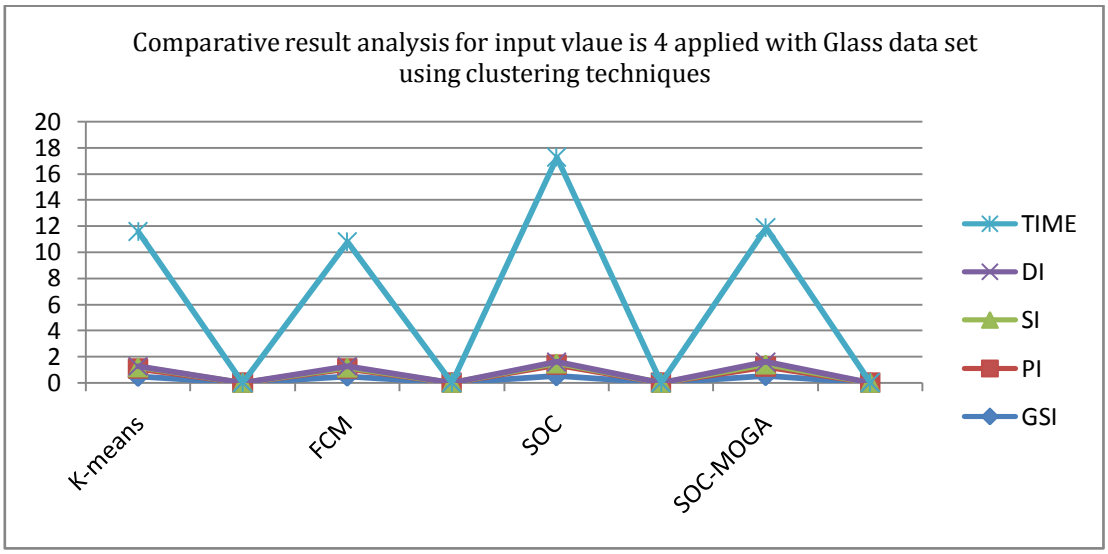


Figure 2: Shows that the comparative result for Glass dataset using clustering techniques with the input value is 4

## VI. CONCLUSION AND FUTURE WORK

In this paper proposed a PSO based two level weighted variable clustering techniques for multi-view data. In this used the PSO inference rule for the selection of important parameter such as eta and lemda, this parameter decides the selection of center point of cluster technique. The automated weighted clustering technique decides the cluster level wise seed and generates cluster according to their features attribute of multi-view data. The experiments also revealed the convergence property of the view weights in Proposed. We compared Proposed with three clustering algorithms on three real-life data sets and the results have shown that the proposed algorithm significantly outperformed the other three clustering algorithms in four evaluation indices. The proposed algorithm is very efficient clustering technique for multi-view data. The PSO algorithm takes more time for the selection of estimated value of lemda. The values of lemda influence the cluster quality during view of data. In future used optimization technique for self-selection of optimal cluster for multi-view data

## 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.