Enhanced the Performance of Self-Optimal Clustering Technique Using Particle Swarm Optimization

[1]Kavita Firke, [2] Dinesh Kumar Sahu

1 M.Tech Scholar, Department of CSE 2 Asso. prof., Department of CSE
Sri Satya Sai College of Engineering, RKDF University, Bhopal M.P., India

ABSTRACT

Self-optimal clustering technique is great advantage over partition clustering technique. the partition clustering technique faced a problem of the generation of cluster and quality validation of generated cluster. The optimal clustering technique automatically decided the number of cluster according to their selection of center point and generation of cluster. In this paper used particle swarm optimization technique for the fitness constraints function for the selection of center point. the particle of swarm optimization gives the dual fitness constraints for the selection of cluster center and quality validation. The modified self-optimal clustering technique implemented in MATLAB software and used reputed data set from UCI. Our experimental result shows that better performance instead of SOC algorithm.

Keywords: -*PSO, SOC, Clustering, EA, SVM***.**

INTRODUCTION

The terms data mining, patent mining, text mining and visualization are employed for the processing of the documents. This chapter will try to give some explanations of the terms and explain why "data mining" was chosen for the title of the study. Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.Clustering is a division of data into groups of similar objects. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. It models data by its clusters. Data modelling puts clustering in a historical perspective rooted in mathematics, statistics, and numerical analysis [16]. From a machine learning perspective clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.Clustering is the subject of active research in several fields such as statistics, pattern recognition, and machine learning. This survey focuses on clustering in data mining. Data mining adds to clustering the complications of very large datasets with very many attributes of different types. This imposes unique computational requirements on relevant clustering algorithms [11]. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems. Some real-life data mining problems involve learning classifiers from imbalanced data, which means that one of the classes called a minority class includes much smaller number of examples than the others classes called as majority classes. Typical such problems are medical diagnosing dangerous illness, analyzing financial risk, detecting oil spills in satellite images, predicting technical equipment failures or information filtering. Class imbalance constitutes a difficulty for most learning algorithms, which are biased toward learning and prediction of the majority classes. As a result, minority examples tend to be misclassified. The information in the minority class is very less as compared to the majority class samples. It is easy to be overlapped by the information of the majority and lead to misclassification. As a result, the performance of the classifier based on balanced data sets is far better than that based on the imbalanced ones. Therefore, the traditional classification approaches and their evaluation criteria are not suitable for the imbalanced data set [7]. The achievements in classifiers based on imbalanced data sets have been presented with different approaches. Sampling method based on the pre-processing of data, which reconstructs the data set artificially to reduce the degree of imbalance. Over-sampling is to increase the number of the minority, but it may lead to over-fitting because of the duplication of data, While,

under-sampling is to cut down the number of the majority class samples. But it may lose information of the majority and decrease the performance of classification. The other method focuses on the algorithm based approach, which introduces certain mechanism to handled the imbalance and make it suitable for the classification on imbalanced data sets. Examples of such techniques are: cost sensitive, support vector machines algorithm (SVM), and some ensemble methods. There are many mechanisms in revising algorithms for imbalanced data mining. For example, the use of adjustment of cost function, the use of different values of weight, the change of probability density. Cost sensitive study algorithm uses the cost of each class to make classification decision.

The rest of paper discuss as in section 2 discuss the Evolutionary Algorithm. In section 3 discuss the Feature Selection. In section 4 discuss proposed Work. In section 5 discuss the experimental result and analysis.Finally discuss conclusion & future work in section 6.

2.EVOLUTIONARY ALGORITHM

EAs utilize the vocabulary obtained from hereditary qualities. They recreate the advancement over an arrangement of eras (cycles inside an iterative procedure) of a populace (set) of applicant arrangements. An applicant arrangement is inside spoken to as a series of qualities and is called chromosome or person. The position of a quality in a chromosome is called locus and all the conceivable qualities for the quality shape the arrangement of alleles of the separate quality. The interior representation (encoding) of a competitor arrangement in a developmental calculation frames the genotype that is prepared by the transformative calculation. Every chromosome relates to an applicant arrangement in the pursuit space of the issue, which speaks to its phenotype. An unraveling capacity is important to make an interpretation of the genotype into phenotype. Transformation and Crossover are two oftentimes utilized administrators alluded to as developmental methodologies [12].

Transformation comprises in an arbitrary irritation of a quality while hybrid goes for trading hereditary data among a few chromosomes, therefore keeps away from neighborhood optima. The chromosome subjected to a hereditary administrator is called parent and the came about chromosome is called posterity. A procedure called choice including some level of haphazardness chooses the people to Recombination or Crossover makes offspring's, chiefly in view of individual legitimacy [16]. The individual legitimacy is assessed utilizing a wellness capacity, which evaluates how the hopeful arrangement befitted being encoded by the chromosome, for the issue being understood. The wellness capacity is detailed in light of the scientific capacity to be advanced. The arrangement returned by a transformative calculation is normally the most fitted chromosome in the last era.

3. 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 predicatively features for the remaining “difficult” classes. As a result, the features that are necessary for discriminating “difficult” classes would be ignored by traditional feature scoring methods [4]. This problem is called the “siren pitfall”. It lessens the quantity of elements, evacuates unessential, repetitive, or loud components, and realizes tangible effects for applications: accelerating an information mining calculation, enhancing learning precision, and prompting to better model conceivability.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. 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.

Feature selection techniques is a type of searching techniques for proposing new feature subset, along with the evaluation measures which scores the different feature subsets. The simplest algorithm is to test each possible subset of features finding the one which minimizes the error rate. This is an exhaustive search of the space and is computationally intractable for all but the smallest of feature sets. The choice of evaluation matric heavily influences the algorithm, and it is these evaluation metrics which distinguish between the three main categories of the feature selection algorithms: wrappers, filters and embedded methods [18].

4. PROPOSED ALGORITHM

Self-optimal clustering well knows method for data clustering. In SOC cluster is suffered from a selection of k number of cluster for level. The selection of optimal number of cluster improves the performance of cluster weighted cluster for multi-category data clustering. Machine learning play big role in pattern recognition and network security. The recognition of pattern faced the series of training process. The training process of clustering technique generates the accuracy performance of classifier and method of pattern recognition. In phase of dataset training imbalance of data arise a problem of minority and majority of class labeling.

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

$$D_k\left(N_{i,R_j}\right) = \frac{Wij\epsilon\,(Cij)}{\sum_i p\,(i,j)}, Ri\epsilon Lk(Ci,Ri) \dots\dots\dots\dots\dots (4.1)$$

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.

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.

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

PROPOSED MODEL

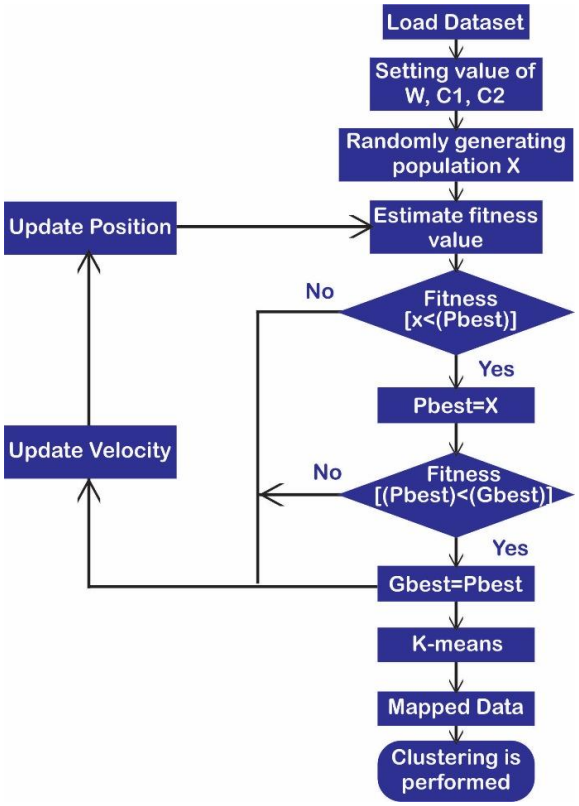

Figure 1: shows that proposed model of self-optimal clustering technique.

5. EXPERIMENTAL RESULT ANALYSIS

For the evaluation of proposed model used matlab software. This dataset contains protein localization sites with multivariate characteristics. The below figures describes the experimental steps followed to process ecoil dataset, where ecoild dataset is taken as input to the MATLAB program and successive steps describes the output created for the input dataset ecoil.

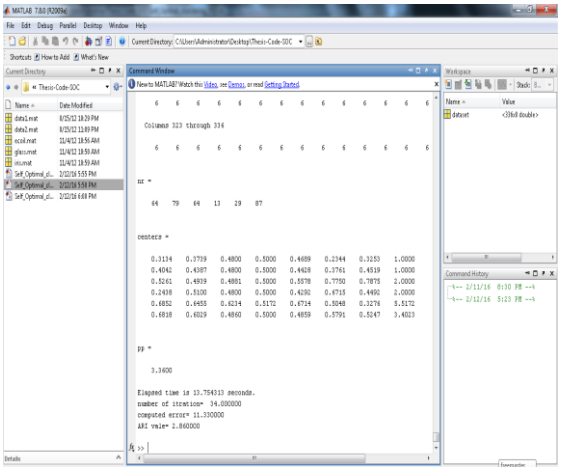

Figure 2: Shows the result value for Ecoil input having K=1 in SOC method.

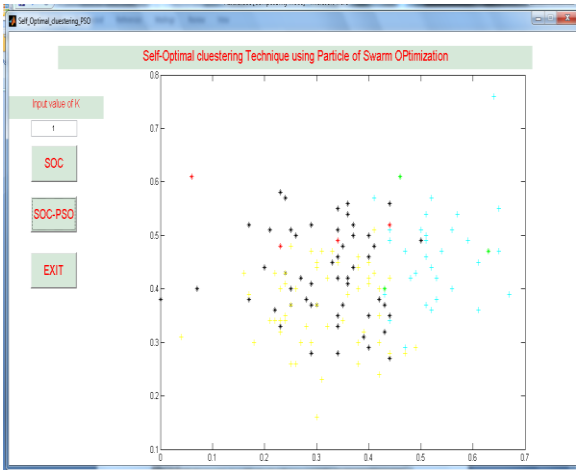

Figure 3: Shows the output image for Ecoil input having K=1 in SOC-PSO method.

For Ecoil dataset having K=1

Value OfK	Method	Elapsed time (seconds)	Number of iteration	Computed error	ARI value
1	SOC	14.354549	34.000000	14.540000	2.860
	SOC-PSO	16.573459	35.000000	9.220000	3.260

For Glass dataset having K=1

Value Of K	Method	Elapsed time (seconds)	Number of iteration	Computed error	ARI value
1	SOC	8.320258	21.000000	6.620000	1.640000
	SOC-PSO	11.195804	22.000000	3.770000	2.040000

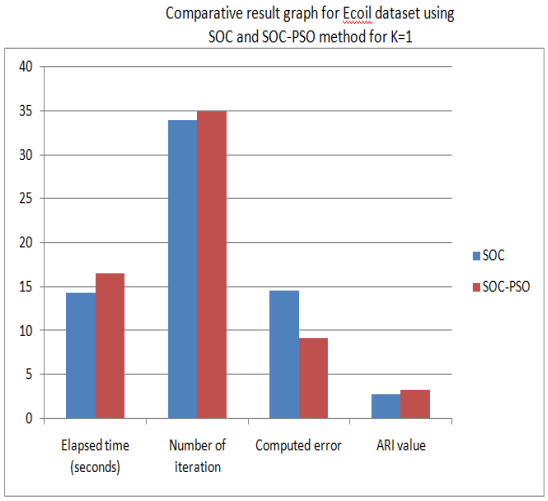

Figure 4: Shows that result for Ecoil dataset using SOC and SOC-PSO for K=1

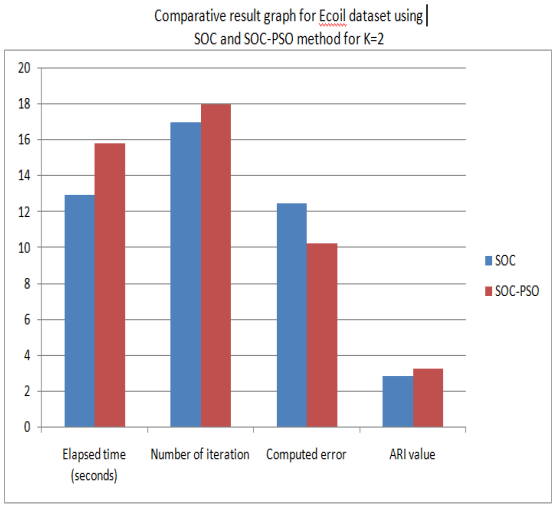

Figure 5: Shows that result for Ecoil dataset using SOC and SOC-PSO for K=2

6.CONCLUSION AND FUTURE WORK

In this paper proposed a PSO based self-optimal clustering techniques for large data. The particle of swarm optimization optimized the value of index and center point of partition cluster.Proposed SOC-PSO clustering algorithm for clustering of large data, Proposed can compute index for views and individual variables simultaneously in the clustering process. With the two types of index, compact views and important variables can be identified and effect of low-quality views and noise variables can be reduced. Therefore, Proposed can obtain better clustering results than SOC clustering algorithms from data. We used four datasets for the experimental process. All dataset obtained from UCI machine learning website. 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] Yuwen Huang "Dynamic Cost-sensitive Ensemble Classification based on Extreme Learning Machine for Mining Imbalanced Massive Data Streams", SERSC, 2015, Pp 333-346.

[2] Benjamin X. Wang and Nathalie Japkowicz "Boosting support vector machines for imbalanced data sets", Springer, 2009, Pp 1-20.

[3] M. Mostafizur Rahman and D. N. Davis "Addressing the Class Imbalance Problem in Medical Datasets", International Journal of Machine Learning and Computing, 2013, Pp 224-228.

[4] R. J. Lyon, J. M. Brooke, J. D. Knowles and B. W. Stappers "A Study on Classification in Imbalanced and Partially-Labelled Data Streams", IEEE, 2013, Pp 1-6.

[5] Ramachandra Rao Kurada, Dr. K Karteeka Pavan, and Dr. AV Dattareya Rao "A PRELIMINARY SURVEY ON OPTIMIZED MULTIOBJECTIVE METAHEURISTIC METHODS FOR DATA CLUSTERING USING EVOLUTIONARY APPROACHES", IJCSIT, 2013, Pp 57-77.

[6] R. J. Lyon, J. M. Brooke, J. D. Knowles and B. W. Stappers "Hellinger Distance Trees for Imbalanced Streams", International Conference on Pattern Recognition, 2014, Pp 1-6.

[7] Yasser Ganjisaffar, Thomas Debeauvais, Sara Javanmardi, Rich Caruana and Cristina Videira Lopes "Distributed Tuning of Machine Learning Algorithms using MapReduce Clusters", ACM, 2011, Pp 1-8.

[8] Volkan Tunali, Turgay Bilgin, and Ali Camurcu "An Improved Clustering Algorithm for TextMining: Multi-Cluster Spherical K-Means", International Arab Journal of Information Technology, 2015, Pp 12-19.

[9] Rahul Malviya, Astt Prof.sushil Tiwari and Prof.S.R.Yadav "A Survey of Modified Support Vector Machine using Particle of Swarm Optimization for Data Classification", Journal of Advanced Computing and Communication Technologies, 2015, Pp 27-32.

[10] A. Fernandez, S. Garcia, and F. Herrera "Addressing the Classification with Imbalanced Data: Open Problems and New Challenges on Class Distribution", Springer, 2012, Pp 1-10.

[11] Shafiq Alam, Gillian Dobbie, Yun Sing Koh, Patricia Riddle and Saeed Ur Rehman "Research on particle swarm optimization based clustering: A systematic review of literature and techniques", Elsevier, 2014, Pp 1-13.

[12] LUO Xin "Chinese Text Classification Based on Particle Swarm Optimization", NCEECE, 2015, Pp 53-58.

[13] Rukshan Batuwita and Vasile Palade "Efficient Resampling Methods for Training Support Vector Machines with Imbalanced Datasets", IEEE, 2010, Pp 1-8.

[14] Spencer Angus Thomas and Yaochu Jin "Reconstructing Biological Gene Regulatory Networks: Where Optimization Meets Big Data", Springer, 2014, Pp 1-15.

[15] Mikel Galar, Alberto Fern´andez, Edurne Barrenechea, Humberto Bustince and Francisco Herrera "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches", IEEE, 2012, Pp 462-484.

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

[17] 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.

[18] RamachandraRaoKurada, K KarteekaPavan, AV DattareyaRao "A preliminary survey on optimized multiobjectivemetaheuristic methods for data clustering using evolutionary approaches" International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, 2013. Pp 57-78.

[19] 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.

[20] RushiLongadge, 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.