IJARCCE

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

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 6, June 2016

Self Optimal Clustering Technique Based on Multi-Objective Genetic Algorithm

Anuradha Paliwal[1], Himanshu Yadav[2], Anurag Jain[3]

Master Student, Dept of Computer Science, Radharaman Institute of Technology and Science, Bhopal, India[1]

Department of Computer Science, Radharaman Institute of Technology and Science, Bhopal, India[2, 3]

Abstract: The self optimal clustering technique is new area of research in data mining. The self optimal clustering technique increases the efficiency and scalability of partition clustering and mountain clustering technique. The concept of self optimal clustering technique used the concept of heuristic function for the selection of cluster index and centre point. In this paper proposed a novel self optimal clustering technique using multi-objective genetic algorithm. The multi-objective genetic algorithm work in two phases in first phase the genetic algorithm work for the selection of center point and merging the cluster index value based on defined fitness constraint value. In second phase of genetic algorithm check the assigned number of value of K for the process of clustering and validated the clustering according to the data sample. The proposed algorithm implemented in MATLAB software and used some reputed dataset form UCI machine learning repository.

Keywords: Data Mining, Clustering, Heuristic Function, MOGA.

I. INTRODUCTION

Clustering well knows technique for the grouping of object on the basis of similarity. The similarity based object collection process is done by partition and fuzzy based clustering technique [1, 2]. The fuzzy and partition based clustering technique suffered a problem of center point selection and cluster content validation [3, 4]. For the validation and purity of clustering technique used threshold based clustering technique. The threshold based clustering technique gives the self optimal clustering process. The self optimal clustering technique estimate how many clusters are generated on the given data. For the process of threshold used various heuristic based optimization algorithms for the selection of threshold function. The heuristic based function used the concept of fitness function; the fitness function decided the selection policy for the generation of cluster. Now a day's various authors used the concept of dual fitness function for the optimization of clustering process. In this paper proposed the self optimal clustering technique for the minimization of the process of cluster validation and reduces the complexity of cluster generation. For the selection of fitness constraints used multi-objective genetic algorithm. The multi-objective genetic algorithm works in dual mode scenario for the generation of cluster. The MOGA derivates function merger the index of intermediate cluster [8]. The intermediate cluster validates the content of cluster. This technique gives better results in terms of cluster validity and time complexity [6].We are able to get all-relevant clusters by reducing the number of redundant clusters. Most of the clusters are demarcated with good performance with this technique. The threshold function defined in IMC is heuristically estimated, always leaving scope for much better optimization of the threshold function and thus having further opportunity in obtaining better quality of clusters. Utilizing this opportunity, we have proposed a self-optimal clustering (SOC) technique with a mathematically optimized threshold function using an interpolation method and compared it with some of the well-known and widely used clustering techniques. It has been shown that the proposed technique is more effective at the optimum number of clusters with better visualized results and well supported by various validity indices as well. The rest of paper organized in the form of section II discusses MOGA algorithm. In section III discuss the proposed method. In section IV discuss the experimental process and finally discuss the conclusion and future work in section V.

II. MOGO

In this section discuss the multi-objective genetic algorithm. The multi-objective genetic algorithm used for the dual selection of fitness constraints for the validation of cluster index and optimality of partition clustering process. The used MOGA process for the weighting of cluster index of scalar multiple objective function. Therefore the search space of multi-objective function is not fixed. A tentative set of optimal solution is preserved in the execution of self optimal clustering technique. Here discuss the process of MOGA for self optimal clustering technique.

1. Selection process

Combined all the fitness constraints function and produces multiple objective functions using the concept of weighting sum of index.

$$f(x) = I1.f(x) + \cdots.+In.f(x) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(1)$$

Where x is a string, f(x) is a combined fitness function, I is the index of optimal cluster

DOI 10.17148/IJARCCE.2016.5612

2. Elite Preserves policy

In process of execution of the MOGA, a tentative set of praetor optimal solutions is strode and update at every generation.

3. Process steps

Step 1 generates an initial population contain N strings where N is the number of string in each population.
Step 2
Estimate the values of the objective function for the generated string. Update the tentative set of solution.
Step3
Estimate the fitness value of each string using the random index. The random index select in terms of pair string for the processing of data.

$$p(x) = \frac{f(x) - fmin(i)}{\sum_{x\in I} f(x) - fmin(i)} \ldots\ldots\ldots\ldots\ldots..(2)$$

Where $fmin(x) = min|f(x), x \in I|$
Step 4
For each selected pair, apply a crossover operation to generate two new strings. N new string is generated by the crossover operation.
Step 4
For each bit value of string generated by the crossover operation apply a mutation operation probability.
Step 5
Randomly remove the N population string from the set of pervious set of N population.
Step 6 if the result oriented solution is obtained the process is terminated.

III.PROPOSED ALGORITHM

The proposed algorithm is combination of two algorithms one is k-means algorithm and other is MOGA (multi-objective genetic algorithm). The MOGA algorithm validates the value of cluster index and validated the generated cluster for the processing of data in terms of cluster index. The proposed algorithm describe in following steps.

Steps 1 Distribution of data sample
Let us consider $\{x^* (i,j) I\ i = 1, 2, \ldots\ldots, n ; j = 1, 2, \ldots, p\}$ is a sample dataset, n is the number of x^* and p is the number of factors of each x^*. $x^*(i,j)$ is the evaluation index j of the i-Th sample. For different quantities of each attribute and different ranges of data, for the formation of cluster data used mapping function using equation (1) and (2)

For the estimation of center point value used these formula

$$x(i,j) = \frac{(x^*(i,j) - x_{min}(j))}{(x_{max}(j) - x_{min}(j))} \qquad (1)$$

For the validation of the cluster index used this formula:

$$x(i,j) = \frac{((x_{max}(j) - x^*(i,j))}{(x_{max}(j) - x_{min}(j))} \qquad (2)$$

Where x_{min} (j)the minimum value of is attributej, and x_{max} (j) is the maximum value of attributej.

Step 2 estimates the value of cluster index $Q(a)$. $\{x(i,j) I\ j = 1, 2 \ldots., p\}$ is distributed into search space based on a genetic algorithm cluster to get index values $z(i)$ through distribuation $a = [a(1), a(2), \ldots.., a(p)]$ as:

$$z(i) = \sum_{j=1}^{p} a(j)x(i,j), \quad i = 1, 2, \ldots\ldots.n \qquad (3)$$

Then, z(i) is the center of cluster, which required validation of mapped cluster space
The evaluation index function of cluster validation is determined by$Q(a)$, shown as:

$$Q(a) = S_z D_z \qquad (4)$$

Where S_z is the standard deviation of $z(i)$; D_z is the fitness value; standard deviation S_z and local density D_z are defined in formula (5):

$$\begin{cases} S_z = \sqrt{\dfrac{\sum_{i=1}^{n}(z(i) - E(z))^2}{(n-1)}} \\ D_z = \displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n}\big(R - r(i,j)\big)\, u(R - r(i,j)) \end{cases} \qquad (5)$$

(1) Defining $d\big(z(k), z(h)\big)$ as the absolute distance between the two cluster index value

$$d\big(z(k), z(h)\big) = \sqrt{\big(z(k) - z(h)\big)\big(z(k) - z(h)\big)}$$
$$= \sqrt{\big(z(k) - z(h)\big)^2}$$

$k = 1, 2, \ldots\ldots\ldots.., N; h = 1, 2, \ldots\ldots\ldots.., N$
$N(n \geq N \geq 2)$ is evaluation level number or the clusters number. And $D_q(q = 1, 2, \ldots\ldots, N)$ is used to describe the intracluster distance of group $G_q(q = 1, 2, \ldots\ldots, N)$,

Step 4 Determining constraints established the self optimal, which gave the constraint conditions est. $\sum_{j=1}^{p} a^2(j) = 1$. but it did not specify a value range.

$$\begin{cases} s.t. \sum_{j=1}^{p} a^2(j) = 1 \\ 1 \geq a(j) \geq 0 \end{cases} \qquad (7)$$

Step 5 cluster evaluation
The evaluation of cluster used formula (3) is used to calculate index values. Sample points which have similar index values are divided into one cluster

IV.EXPERIMENTAL ANALYSIS

In this paper perform experimental process of modified SOC with MOGA. The proposed method implements in MATLAB 7.14.0 and tested with very reputed data set from UCI machine learning research center.

To evaluate these performance parameters I have used four datasets from UCI machine learning repository [10] namely Iris, Glass identification, Diabetes and Ecoil data set. Out of these four dataset, two are small dataset namely Iris and Glass identification dataset; and remaining two are large datasets namely Diabetes and Ecoil data set

DOI 10.17148/IJARCCE.2016.5612

PERFORMANCE TABLE OF ALL DATA SET FOR CLUSTERING TECHNIQUES

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-MOGA	3.800	2.387	0.748	0.718	29.144

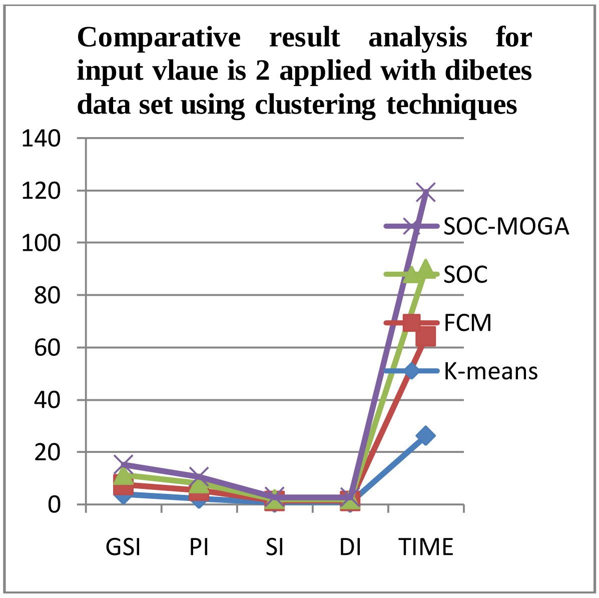

Figure 1: Shows that the comparative result for diabetes dataset using clustering techniques with the input value is 2.

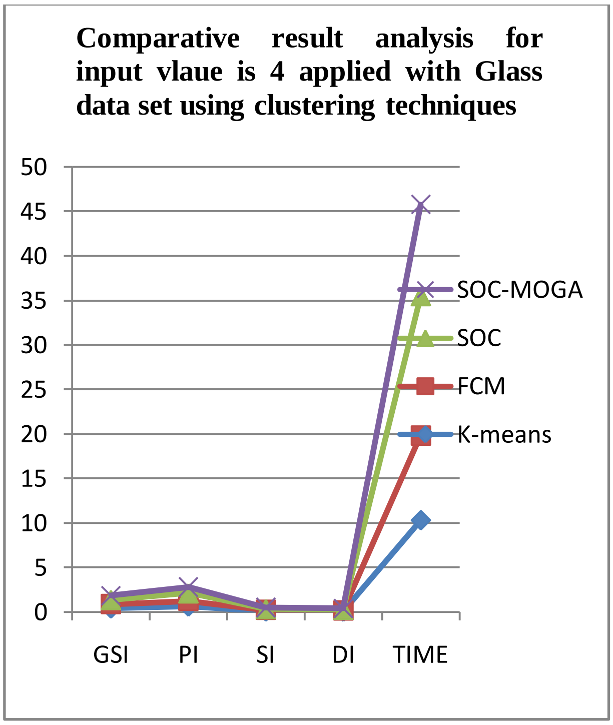

Figure 2: Shows that the comparative result for Glass dataset using clustering techniques with the input value is 4.

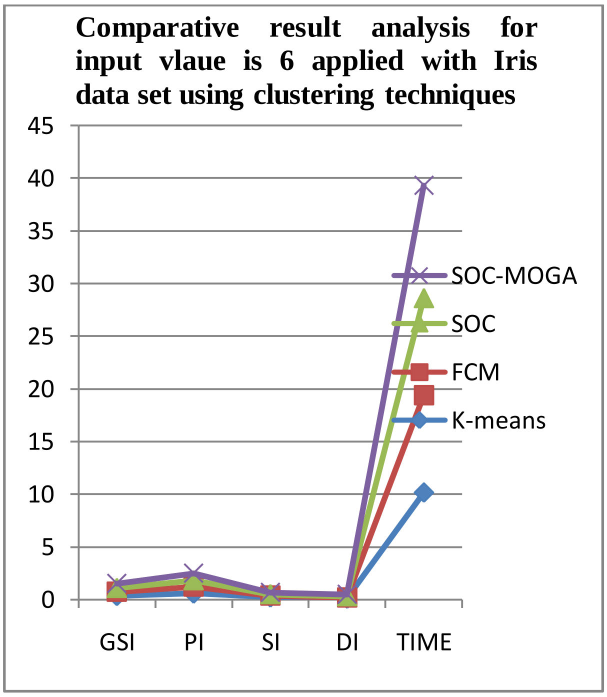

Figure 3: Shows that the comparative result for Iris dataset using clustering techniques with the input value is 6.

V. CONCLUSION & FUTURE SCOPE

In this paper proposed a novel method of self optimal clustering technique based on multi-objective genetic algorithm. Here multi-objective genetic algorithm creates the dual search space for the index and center point. The dual fitness constraints create the scalar space of fitness constraints. The designed fitness constraint minimized the index ratio value and increases the quality of clustering process. The proposed clustering technique improved the performance of self-optimal clustering technique. In future reduces the selection set process of multi-objective genetic algorithm for the better performance of self-optimal clustering technique.

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.

DOI 10.17148/IJARCCE.2016.5612

[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.

DOI 10.17148/IJARCCE.2016.5612