Iranian Journal of Optimization
Volume 10, Issue 1, 2018, 51-58
Research Paper
Online version is available on: www.ijo.iaurasht.ac.ir

Islamic Azad University
Rasht Branch
E-ISSN:2008-5427

Improved Automatic Clustering Using a Multi-Objective Evolutionary Algorithm with New Validity Measure and Application to Credit Scoring

Majid Mohammadi Rad[1*] and Mahdi Afzali[2]

[1] Department of Computer and Information Technology, Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran
[2] Faculty of Computer Engineering, Islamic Azad University, Zanjan Branch, Zanjan,Iran

Received: 29 June 2017
Accepted: 22 September 2017

Keywords:
Clustering
Data Mining
Evolution Algorithm
Credit Score
Clustering Validity Meassure

Abstract

In data mining, clustering is one of the important issues for separation and classification with groups like unsupervised data. In this paper, an attempt has been made to improve and optimize the application of clustering heuristic methods such as Genetic, PSO algorithm, Artificial bee colony algorithm, Harmony Search algorithm and Differential Evolution on the unlabeled data of an Iranian bank with the credit scoring approach. A survey was also used to measure the clustering validity index which resulted in a new validity index. Finally, the results were compared to identify the best algorithm and validity measure (Das & Konar, 2009).

***Correspondence E_mail**: mrad@aut.ac.ir

INTRODUCTION

In this paper, clustering of a data set is viewed as an optimization. Evolutionary algorithms such as Genetic, PSO algorithm, artificial bee colony algorithm and Harmony Search algorithm have been utilized in order to solve a problem (Holland, 1975).

In addition the automatic decision of the optimal number of clusters in unlabeled data set beside applying evolutionary algorithm for automatic reclustering problem, and effect a clustering validity measure to provide global max/min of classes (Chou et al., 2004).A new validity index was proposed to use and consequently the results were meticulously compared to identify the best algorithm and validity indexes.

This study is divided into eight sections: Section 2 describes the concepts of Credit Scoring and algorithm. Section 3 describes the previous researches about clustering and credit scoring. Section 4 is about problem explanation. Section 5 describes the New Validity Measure and Section 6 describes data set and detected rules for credit score and

Concepts of clustering and evaluation of the new validity measure is discussed in section 7 and Section 8 is devoted to results Section.

ASIC CONCEPT

K-Nearest Neighbor

KNN is one of the non-parametric classifiers which is based on learning by similarity. In this method, the space pattern for the K nearest neighbor is explored for each new observation, that is the closest to the new observation in term of distance from the explanatory variables (Paredes & Vidal, 2000) and (Hand & Vinciotti, 2003) and (Islam et al., 2007) and (Marinakis et al., 2008) and (Li, 2009).

Clustering

In clustering, similar data are grouped into the same cluster (Das et al., 2008). Partition clustering algorithms try to separate the data set into a set of disjoint clusters and also attempt to optimize and improve certain criteria (Das et al., 2008).

Cluster validity indexes match with the statistical mathematical function which is used to evaluate the clustering result beside concerning the appearance of the clustering.

(1) Compression: as the highest similarity and affinity.

(2) Dissociation: as the lowest similarity among members of the clusters (Das et al., 2008).

The maximum or minimum values of these indexes indicate the suitable partitions (Das & Konar, 2009). Due to their optimizing character, the cluster validity indexes are used efficiently in association with any optimization algorithm such as GA, PSO, HS, ABC,etc (Das et al., 2008).

PREVIOUS RESEARCHESP

Several methods in the field of clustering, the estimated credit risks, have been applied (Harrell & Lee, 1985). Intelligent methods, SVMs and neural network are used frequently and have classified the accuracy rate appropriately (Desai et al., 1997); (Huang et al., 2007).Neural network ensemble strategies include bagging and boosting for financial decision applications which have been studied and shown better accuracy rate(West et al., 2005); (Sadatrasoul et al., 2015) . Some studies have shown the superiority of the decision trees, neural networks and other intelligent methods to statistical methods (Crook et al., 2007); (Das & Konar, 2009); (Sadatrasoul et al., 2015).

PROBLEM DEFINITION

Data mining technique can contribute to identifying algorithms and hidden knowledge of information in data sets, where clustering is one of the most significant and applicable concepts (Das et al., 2008).

Identifying the best validity index with the use of algorithm for the best automatic clustering with K-means is the purpose of the present study (Srikrishnar et al).

In order to enhance the process of clustering and obtain accurate results, the new validity index, which is introduced below, has been used. This index is a combination of CS index and DB index and is set with a parameter value. The present article introduces the above mentioned index in line with the improvement of the evaluation criteria of.

Counting the number of clusters is applied

manually and obtaining the cost of the applied function and comparing it with the automatic clustering method is evaluated. Details of the findings of are demonstrated in the attached tables.

NEW VALIDITY MEASURE

(1)DB Index

This amount is a planning of the proportion of the sum of within-cluster between cluster separation. The smallest DB index indicates a valid optimal partition (Das et al., 2008).

To present the state in Mathematical model, the following equations can be utilized:

X={x1,x2,...,xn} xi∈ R
M={m1,m2,...,mk} Cluster Center
K=Number of Cluster

$$\S i,q=\sqrt[q]{\frac{1}{Ni}*\Sigma d(x,mi)^q} \quad (1)$$

$$di,j,t=\sqrt[t]{\Sigma|mip-mjp|^t}\equiv\|mi-mj\|_t \quad (2)$$

$$Hi,q,t=\max\frac{di,q+dj,q}{di,j,t} \quad (3)$$

$$DB(k)=\frac{1}{k}\Sigma Hi,q,t \quad (4)$$

(2)CS Measure

Evaluating the validity of clustering and the centroid of a cluster were computed. The CS measure is defined as (Das et al., 2008).

$$CS(k)=\frac{\Sigma[\frac{1}{Ni}\Sigma\max\{d(\overline{Xi},\overline{Xq})\}]}{\Sigma[\min\{d(mi,mq)\}]} \quad (5)$$

(3)NEWDBCSmeasure

This validity index is a combination of DB and CS and proposed for clustering and the centroid of cluster computation.

$$NEWDB=(NEWDBparam\ \ eter*(\frac{1}{k}\Sigma\max\frac{di,q+dj,q}{di,j,t})) \quad (6)$$

$$NEWCS=(NEWCSparam\ \ eter*(\frac{\Sigma[\frac{1}{Ni}\Sigma\max\{d(\overline{Xi},\overline{Xq})\}]}{\Sigma[\min\{d(mi,mq)\}]}) \quad (7)$$

$$NEWDBCS=(\frac{(((NEWDBparam\ eter*(\frac{1}{k}\Sigma\max\frac{di,q+dj,q}{di,j,t})))+(NEWCS))}{\sum(NEWDBparam\ eter,NEWCSparam\ eter)}) \quad (8)$$

DATA SET USED AND ALGORITHM PARAMETRS

The available data are received from an Iranian bank with 610 records,540 records remain after clearance and this number of records are put in a 12*540 matrix which is measured at the input of Genetic, PSO algorithm and Artificial Bee Colony Algorithm and Harmony Search Algorithm with specified parameters according to Table 1 and 2 and 3 and 4.

Afterwards, by reviewing the clusters and data inside them accurately, some rules are detected and applied to the data, some of which are mentioned below:

1. If the amount of debt to income is less than 12, then the customer will be well-off.
2. If the amount of debt to income is higher than 24, then the client will be uncreditworthy.
3. If the job experience is less than 14 years old, the amount of credit is greater than 4, and the amount of debt to his credit is less than 22, then the client is uncreditworthy.
4. If the amount of credit to the debt is below 2, the age of the person is under 37, the other deviations are above 3 and less than 12, the qualification is a bachelor's degree, and if the residence is over 4 years, then the client will be well-off.

Table 1: Genetic Algorithm Parameters For The Clustering

K max=25, K min=2	Number of clusters
MaxIt=200 Maximum	Number of Iterations
pc=0. 8	Crossover Percentage
pm=0. 3	Mutation Percentage
gamma=0. 05	gamma
mu=0. 02	Mutation Rate
beta=8	Selection Pressure

Table 2: Particle Swarm Optimization Algorithm Parameters For The Clustering

K max=25, K min=2	Number of clusters
MaxIt=200 Maximum	Number of Iterations
phi1=2. 05, phi2=2.05,phi=4.1000	Constriction Coe_cients
W=chi=0.7298	Interia Weight
c1=chi*phi1 =1. 4962	Personal Learning Coe_cient
c2=chi*phi2 =1. 4962	Global Learning Coe_cient

Table 3: HS Algorithm Parameters For The Clustering

K max=25, K min=2	Number of clusters
MaxIt=200 Maximum	Number of Iterations
Hms=20	Harmony Memory Size
HMCR=0.2	Harmony Memory Consideration Rate
PAR=0.1	Pitch Adjustment Rate
FWdamp=0.995	Fret Width Damp Ratio

Table 4:DE Algorithm Parameters For The Clustering

K max=25, K min=2	Number of clusters
MaxIt=200	Maximum Number of Iterations
betamin=0.2	Lower Bound of Scaling Factor
betamax=0.8	Upper Bound of Scaling Facto
PCR=0.2	Crossover Probability

EVALUATION

The optimal number of the application of clusters on Heuristic methods such as Genetic, PSO algorithm and Artificial bee colony algorithm and Harmony Search algorithm was automatically determined in an Iranian bank unlabeled data set and affect the clustering validity index to provide global maximum/minimum of classes in the data set and then the results were compared to identify the best algorithm. The two dimensional plots from MATLAB are as follows in fig. 1, 2, 3, 4, 5, 6, 7, 8. According to the Table 5, 6, 7 each algorithm calculated the result by MATLAB for the best cost in first and last iteration described.

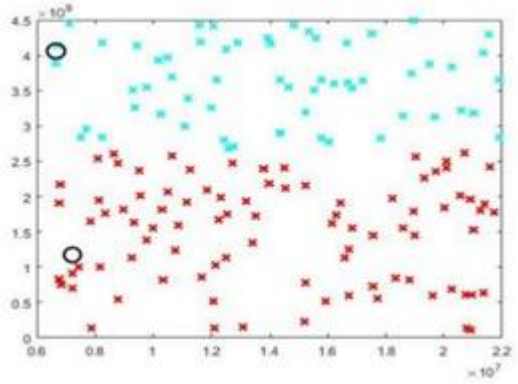

Fig. 1. ABC algorithm with DB

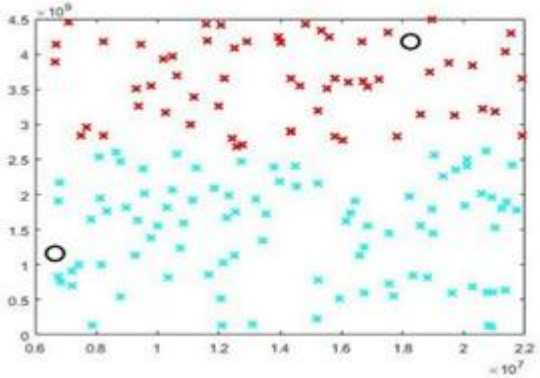

Fig. 2. DE algorithm with D

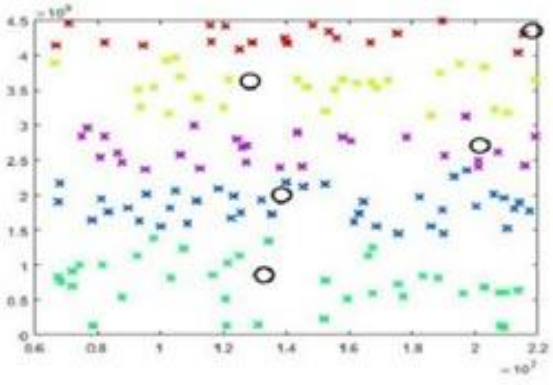

Fig.3. HS algorithm with DB

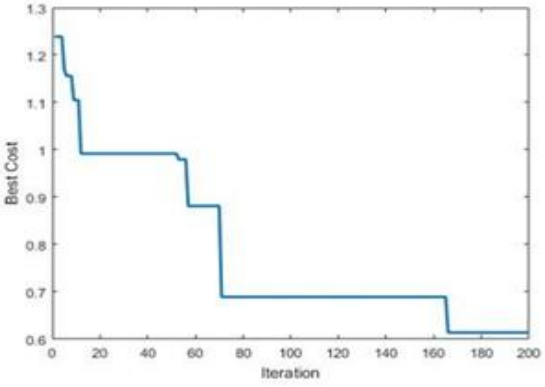

Fig. 4. HS algorithm Best Cost

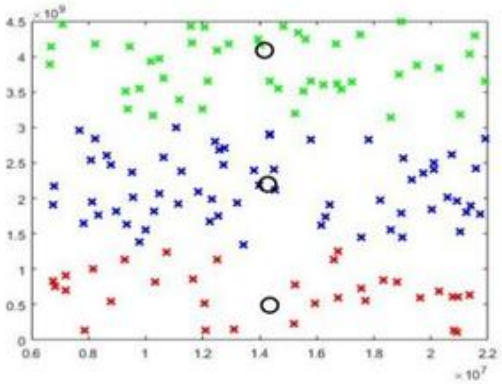

Fig.5. GA algorithm with DB

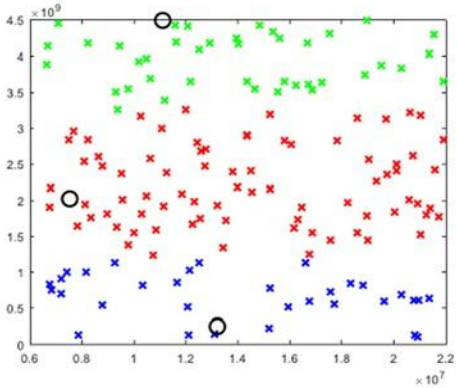

Fig. 6. GA algorithm with NEWDBCS

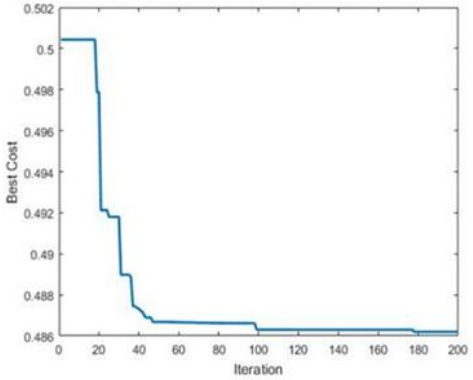

Fig.7. GA algorithm Best Cost

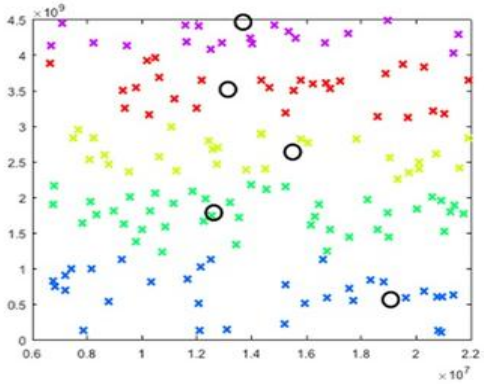

Fig. 8. PSO algorithm with NEWDBCS

Table 5: Result for Best Cost value and optimum cluster number with algorithms

Algorithm	GA			PSO		
Validity indexes	DB	CS	NEW	DB	CS	NEW
Validity indexes Value	0.5248	0.6887	0.6652	0.5227	0.5540	0.7193
First Iteration Best Cost	0.5312	0.69736	0.79561	1.56156	0.755265	1.2561
last iteration Best Cost	0.52478	0.68579	0.66522	0.52267	0.55397	0.71938
Number of Clusters	3	6	3	3	15	5

Table 6: Result for Best Cost value and optimum cluster number with algorithms

Algorithm	HS			ABC		
Validity indexes	DB	CS	NEW	DB	CS	NEW
Validity indexes Value	0.7426	0.5955	0.5777	0.5305	0.5661	0.8781
First Iteration Best Cost	1.1031	0.67924	0.71194	1.1183	0.72533	1.3819
last iteration Best Cost	0.74262	0.59552	0.57773	0.53051	0.56613	0.87816
Number of Clusters	5	7	7	2	11	3

Table 7: Result for Best Cost value and optimum cluster number with algorithms

Algorithm	DE		
Validity indexes	DB	CS	NEWDBCS
Validity indexes Value	0.5326	0.5565	0.5625
First Iteration Best Cost	1.1732	0.72533	0.71262
last iteration Best Cost	0.53259	0.55651	0.56258
Number of Clusters	2	9	8

CONCLUSION

In this paper, unlabeled data obtained from an Iranian bank based on k-means method, were clustered using meta-heuristic methods. The optimal number of clusters was automatically calculated with the use of genetic and PSO algorithm with the application of validity indexes and proposing a NEWDBCS validity index and use in code. It was concluded that employing evolution algorithm methods can improve clustering operation. The result of best cost in last iteration reveals the fact that Genetic Algorithm with NEWDBCS validity index can serve as the most efficient solution for the automatic clustering in this data set.It is also possible to improve the estimation of credit risk of the customers by identifying the extracted rules in each cluster.

REFERENCES

Ben-David, A. (2008). Rule effectiveness in rule-based systems: A credit scoring case study. *Expert Systems with Applications, 34*(4), 2783-2788.

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods, 3*(1), 1-27.

Chou, C. H., Su, M. C., & Lai, E. (2004). A new cluster validity measure and its application to image compression. P*attern Analysis and Applications, 7*(2), 205-220.

Crook, J. N., Edelman, D. B., & Thomas, L. C. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational*

Research, 183(3), 1447-1465.

Das, S., Abraham, A., & Konar, A. (2008). Automatic clustering using an improved differential evolution algorithm. I*EEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans, 38*(1), 218-237.

Das, S., & Konar, A. (2009). Automatic image pixel clustering with an improved differential evolution. *Applied Soft Computing, 9*(1), 226-236.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence,* (2), 224-227.

Desai, V. S., Conway, D. G., Crook, J. N., & Overstreet Jr, G. A. (1997). Credit-scoring models in the credit-union environment using neural networks and genetic algorithms. *IMA Journal of Management Mathematics, 8*(4), 323-346.

Dunn, J. C. (1974). Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics, 4*(1), 95-104.

Hand, D. J., & Vinciotti, V. (2003). Choosing k for two-class nearest neighbour classifiers with unbalanced classes. *Pattern recognition letters, 24*(9), 1555-1562.

Harrell, F. E., & Lee, K. L. (1985). A comparison of the discrimination of discriminant analysis and logistic regression under multivariate normality. *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences', North-Holland, New York, United States*, 333-343.

Holland, J. H. (1975). Adaption in natural and artificial systems. *Ann Arbor MI: The University of Michigan Press.*

Huang, C. L., Chen, M. C., & Wang, C. J. (2007). Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications, 33*(4), 847-856.

Islam, M. J., Wu, Q. J., Ahmadi, M., & Sid-Ahmed, M. A. (2007, November). Investigating the performance of naive-bayes classifiers and k-nearest neighbor classifiers. In *Convergence Information Technology, 2007. International Conference on* (pp. 1541-1546). IEEE.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. A*CM computing surveys (CSUR), 31*(3), 264-323.

Keramati, A., & Yousefi, N. (2011, January). A proposed classification of data mining techniques in credit scoring. In Proc. *2011 Int. Conf. on Industrial Engineering and Operations Management Kuala Lumpur, Malaysia.*

Kettani, O., Ramdani, F., & Tadili, B. (2015). AK-means: an automatic clustering algorithm based on K-means. *Journal of Advanced Computer Science & Technology, 4*(2), 231-236.

Kuo, R., & Zulvia, F. (2013). Automatic clustering using an improved particle swarm optimization. *Journal of Industrial and Intelligent Information, 1*(1).

Lahsasna, A., Ainon, R. N., & Teh, Y. W. (2010). Credit Scoring Models Using Soft Computing Methods: A Survey. *Int. Arab J. Inf. Technol., 7*(2), 115-123.

Li, F. C. (2009, August). The hybrid credit scoring strategies based on knn classifier. In *Fuzzy Systems and Knowledge Discovery, 2009. FSKD'09. Sixth International Conference on* (Vol. 1, pp. 330-334). IEEE..

Marinakis, Y., Marinaki, M., Doumpos, M., Matsatsinis, N., & Zopounidis, C. (2008). Optimization of nearest neighbor classifiers via metaheuristic algorithms for credit risk assessment. J*ournal of Global Optimization, 42*(2), 279-293.

Pakhira, M. K., Bandyopadhyay, S., & Maulik, U. (2004). Validity index for crisp and fuzzy clusters. *Pattern recognition, 37*(3), 487-501.

Paredes, R., & Vidal, E. (2000). A class-dependent weighted dissimilarity measure for nearest neighbor classification problems. *Pattern Recognition Letters, 21*(12), 1027-1036.

Sabzevari, H., Soleymani, M., & Noorbakhsh, E. (2007). A comparison between statistical and data mining methods for credit scoring in case of limited available data. *In Proceedings of the 3rd CRC Credit Scoring Conference, Edinburgh, UK.*

Sadatrasoul, S., Gholamian, M., & Shahanaghi, K. (2015). Combination of Feature Selection and Optimized Fuzzy Apriori Rules: The Case of Credit Scoring. *International Arab Journal of Information Technology (IAJIT), 12*(2).

Srikrishna, A., Srinivas, V. S., & Jetson, V. R. A naive Fuzzy Clustering Method for Pixel Segmentation by using Differential Evolution.

Tsai, C. F., & Wu, J. W. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications, 34*(4), 2639-2649.

Van Gestel, T., & Baesens, B. (2009). *Credit Risk Management: Basic concepts: Financial risk*

*compo*n*ents, Rating analysis, models, economic and regulatory capital.* Oxford University Press.

West, D., Dellana, S., & Qian, J. (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research, 32(*10), 2543-2559.

Chou, C. H., Su, M. C., & Lai, E. (2004). A new cluster validity measure and its application to image compression. *Pattern Analysis and Applications, 7*(2), 205-220.

Raposo, C., Antunes, C. H., & Barreto, J. P. (2014, June). Automatic Clustering using a Genetic Algorithm with New Solution Encoding and Operators. *In International Conference on Computational Science and Its Applications* (pp. 92-103). Springer, Cham.