ACSIJ Advances in Computer Science: an International Journal, Vol. 5, Issue 3, No.21 , May 2016
ISSN : 2322-5157
www.ACSIJ.org

ACSIJ
WWW.ACSIJ.ORG

Improving OPTICS Algorithm with Imperialist Competitive Algorithm: Choosing Automatically Best Parameters

Mahdi Ghorbani[1], Dr. Mahdi Esmaeili[*2], Maryam Alizadeh[3]

[1] Department of Computer science, Kashan branch
Islamic Azad University Kashan, Iran
m.ghorbani@iaukashan.ac.ir

[2] Department of Computer science, Kashan branch
Islamic Azad University Kashan, Iran
m.esmaeili@iaukashan.ac.ir

[3] Department of Computer science, Kashan branch
Islamic Azad University Kashan, Iran
alizadeh.maryam68@gmail.com

Abstract

Clustering based on similarity is one of the most important stages in data analysis and a beneficial tool for data mining. There are a wide range of data from which a clear pattern cannot be driven using common clustering methods, data with unusual shapes. Scientists introduced density-based clustering algorithms to resolve this issue and enable clustering for this kind of data.

Among all density-based clustering algorithms, Ordering points to identify the clustering structure (OPTICS) is an algorithm for finding density-based clusters in spatial data. Density-based methods describe clusters as dense areas among diffused ones. An informal definition of cluster is: "For each intra-cluster point, neighborhood with specific radius ε, must contains minimum number of point's μ." It's very difficult to specify these two parameters for any types of data and needs great effort setting up them to achieve desired results.

The main goal of this research is to use meta-heuristic methods especially Imperialist Competitive Algorithm (ICA) to precise estimation of these parameters (ε, μ) so that we can apply them to OPTICS Algorithm to achieve accurate and high quality clusters for any data sets. In order to evaluation of mentioned method, we utilized data sets specialized for classification and removed class label. Comparing our results with original one, proved that our method is produced a precise class label.

Keywords: *Data mining; Density-based Clustering; OPTICS; Meta-heuristics Algorithm; Imperialist Competitive Algorithm.*

1. Introduction

Today the use of data mining techniques is growing because companies and individuals do not have Comprehensive knowledge about every context and achieving effective data management. One of the significant data mining techniques is clustering data as isolated and meaningful sets. [1] When the data which are supposed to be clustered are not normal, we cannot use normal clustering method such as partition based. In such cases density-based methods are used. [2] The principal of these methods is that clusters are dense regions of data diffused ones. In other words, we focused on density.

OPTICS was presented by Mihael Ankerst, et al.[3] Its basic idea is similar to DBSCAN,[4] but it addresses one of DBSCAN's major weaknesses: the problem of detecting meaningful clusters in data of varying density. In order to do so, the points of the database are (linearly) ordered such that points which are spatially closest become neighbors in the ordering. Additionally, a special distance is stored for each point that represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster. This is represented as a dendrogram.

This algorithm is a method for density-based clustering of information which uses two input parameters, Eps (the radius of neighborhood) and MinPts (the minimum points required to constitute a cluster) which in this paper will be known as ε and μ respectively. Every data mining task has the problem of parameters because they impact on the algorithm in specific ways. For OPTICS these parameters are ε and μ. It's very difficult for users to specify accurate values for them. Mostly Statistical and classical methods or both are used to specify these parameters which have high computational complexity and are not very accurate. Recently many researches introduced methods to specify these two parameters.

This paper is organized as follow: In Section 2 we explain OPTICS algorithm and define some terminology, In Section 3 we explain Imperialist Competitive Algorithm (ICA) and describe its structure, In Section 4 we introduce our proposed method (ST-OPTICS) to improve the performance of OPTICS algorithm, In Section 5 we evaluate our method using different data sets and Finally, in Section 6 we represent the conclusions and some suggestions for future works.

2. OPTICS: basic idea

A point p is a core point if at least $MinPts$ points are found within its ε neighborhood $N_\varepsilon(p)$. Contrary to DBSCAN, OPTICS also considers points that are part of a more densely packed cluster, so each point is assigned a core distance that describes the distance to the $MinPts^{\text{th}}$ closest point:

$$Reach-dist_{\varepsilon.MinPts}(o.p) = \begin{cases} UNDEFINED & -\ if|N\varepsilon(p)| < MinPts \\ MinPts-th\ \ smallest\ distance\ \ to\ N\varepsilon(p) & -\ otherwise \end{cases} \quad (1)$$

The reachability-distance of another point o from a point p is either the distance between o and p, or the core distance of p, whichever is bigger:

$$Core-dist_{\varepsilon;MinPts}(p) = \begin{cases} UNDEFINED & -\ if|N\varepsilon(p)| < MinPts \\ \max(Core-dist_{\varepsilon;MinPts}(p).dist(p.o)) & -\ otherwise \end{cases} \quad (2)$$

If p and o are nearest neighbors, this is the $\varepsilon' < \varepsilon$ we need to assume in order to have p and o belong to the same cluster. Both the core-distance and the reachability-distance are undefined if no sufficiently dense cluster (w.r.t.ε) is available. Given a sufficiently large ε, this will never happen, but then every ε-neighborhood query will return the entire database, resulting in $O(N^2)$ runtime. Hence, the ε parameter is required to cut off the density of clusters that is no longer considered to be interesting and to speed up the algorithm this way.

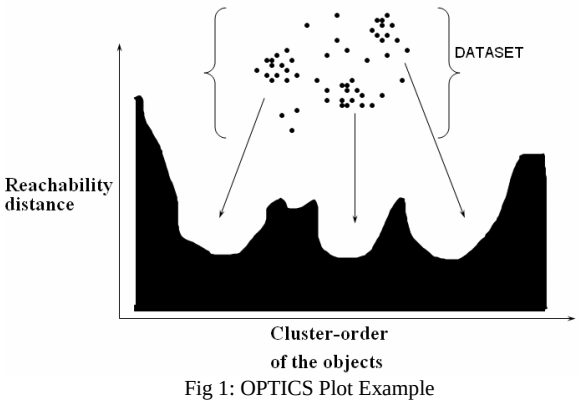

Fig 1: OPTICS Plot Example

2.1 Time and volume complexity function

This algorithm probably meets each points several times (for example it considers which clusters a single point can fit in). In experimental implements, maximum complexity has been met, because of calling Update function to find neighborhood for a single point. The OPTICS do the same process for every point. While indexing structure executes at $O(\log n)$ and value for ε is defined correctly, overall average runtime will be $O(n \log n)$. When the indexing structure is not used data isn't normal (e.g. all points with a distance less than ε) complexity will be $O(n^2)$ at the worst case. The distance matrix can be calculated once with the size of $n^2 - n/2$ and the results can be stored to save time of recalculation but in this case the memory with the size of $O(n^2)$ will be needed. Without using matrix, the memory required to implement OPTICS will be $O(n)$.

2.2 Advantages of method

- quick for small data
- find arbitrary and sphere shaped clusters
- specify noises and robust to outliers
- doesn't need to no number of clusters
- only need two parameters to execute
- Ordering of the points is important only at borders

2.3 Disadvantages of method

- OPTICS isn't completely deterministic at borders (points may feet into two clusters at borders)
- Euclidean distance metric is not applicable for large scale data
- unpredictable facing changes of ε and μ
- not easy for user to set ε and μ

3. Imperialist Competitive Algorithm (ICA)

Inspired by the nature optimization algorithms, have succeed among other classic methods as intelligent optimization methods. Some of the most famous methods are Genetic Algorithms (GA) [5-7] (Inspired by biological evolution of human and other species), Ant Colony Optimization (ACO) [8] (based on optimized movement of ants) and Simulated Annealing (SA) [9] (inspired by annealing process for metal logy). These methods are used to resolve optimization issues in different fields such as determination of optimized path for automatic agents, designing optimized controllers for industry, resolving queue problems and clustering.
ICA is one of the relatively new meta-heuristics optimization algorithms which propose a method to resolve optimization by mathematical modeling of socio-politically evolution process. [10] Same as all algorithms in this category, ICA provides initial population and evaluates them by Eq.(3). This population is known as "Chromosome" in GA, "Particle" in Particle Swarm Optimization (PSO) and "country" in ICA. Basic principles of this algorithm are assimilation, imperialist competition and revolution. Simulating social, economic and political evolution of countries and providing operators as algorithms, ICA helps us to resolve complicated optimization. In fact, this algorithm constructs

empires based on countries, calculates costs by Eq.(4) and finally try's to reach optimum result by a recursive process and optimizing the population gradually.

$$country = [p_1.p_2.p_3.\dots.p_{N_{imp}}]$$
$$cost = f(country) = f\left(p_1.p_2.p_3.\dots.p_{N_{imp}}\right)$$
$$C_n = c_n - max_i\{c_i\}$$
$$P_n = \left|\frac{C_n}{\sum_{i=1}^{N_{imp}} C_i}\right| \quad (3)$$

$$TC_n = Cost(imperialist_n) + \zeta.mean\{Cost(colonies\ of\ empire_n)\} \quad (4)$$

3.1 Assimilation: Moving Colonies toward the Imperialist

According to the algorithm, countries are divided to imperialists and colonies. Considering its power, every imperialist absorbs some of colonies and take them under control. Assimilation is one of the main two principals of this algorithm. Studying the history of grate? imperialists like France and England, they usually tried to wipe out traditions and cultures of colonies by some methods such as constituting schools which uses their languages. This process represented in the algorithm by moving colonies of an empire based on a special equation. Fig.2 and Fig.3 show this movement and variables are defined by Eq.(5), where β is a number greater than 1 and d is the distance between the colony and the imperialist state. Setting $\beta > 1$ causes colonies to get closer to the imperialist state, γ is a parameter that adjusts the deviation from the original direction. Nevertheless, the values of β and γ are arbitrary, in most of implementations setting about 2 for β and about $\pi/4$ (Rad) for γ results in good convergence of countries to the global minimum.

$$x \sim U(0.\beta * d). \quad \theta \sim U(-\gamma.\gamma) \quad (5)$$

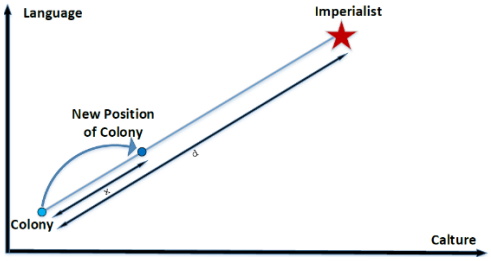

Fig 2: Movement of colonies toward their relevant imperialist

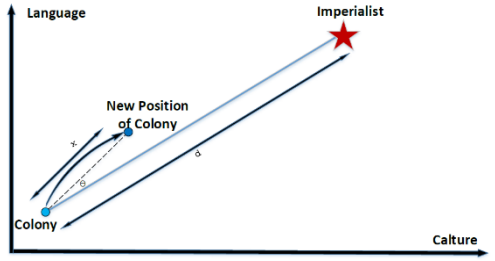

Fig 3: Movement of colonies toward their relevant imperialist in a randomly deviated direction

3.2 Imperialist competition

Imperialist competition is the other important issue of this algorithm. Though competition weak empires gradually lost their power and eventually will be eliminated. This competition leads to a state in which single empire rules the world. This state happens when algorithm reaches optimum solution and stops. Eq.(6) shows calculation method of this process and imperialist competition diagram is shown in Fig.4.

$$P = \left|\frac{N.T.C_n}{\sum_{i=1}^{N_{imp}} N.T.C_i}\right|$$
$$P = [P_{p_1}.P_{p_2}.\dots.P_{p_{N_{imp}}}]$$
$$R = [r_1.r_2.\dots.r_{N_{imp}}].where\ r_i \approx U(0,1)\ and\ 1 \le i \le N_{imp}$$
$$D = P - R = [D_1.D_2.\dots.D_{N_{imp}}] = [P_{p_1} - r_1.\dots.P_{p_{N_{imp}}} - r_{N_{imp}}] \quad (6)$$

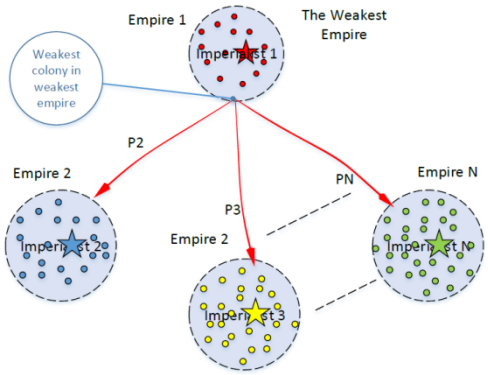

Fig 4: imperialist competition diagram

3.3 Revolution

Revolution causes radical social and political changes in a country. In ICA, revolution is modeled with random movement of a colony to a new position. Revolution saves movements from trapping in local optimums and in some cases improves the position of the country and moves it to a better area. This action is shown in Fig.5.

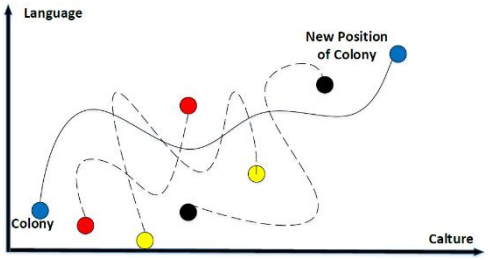

Fig 5: Revolution; radical change in socio-political characteristics of a country

4. Proposed Method: Self-tuning OPTICS (ST-OPTICS)

The principal of this research is to specify a random value for μ during initial state of algorithm and calculate ε using some equations which will be introduced as follow. The idea is to initial the population after setting the parameters of ICA and the position of every country will be considered as value of μ afterwards. These positions will be randomly assigned during the initialization considering maximum and minimum values. The maximum is an optimistic guess because an accurate value is not needed, for example we can apply 30 for every type of data set, but to determine the minimum we must follow some rules which are:

- μ should be greater than 1, otherwise the algorithm will consider every point in space as a cluster
- We can consider the value of μ relative to dimensions (attributes of data set) which means $\mu > D$
- By setting μ to 2 we will have the same hierarchical clustering with the single link metric, with the dendrogram cut at height ε.

Therefore, μ starts at 3 as the minimum then increases. However, larger values are usually better for data sets with noise and will lead to more significant clusters. The larger the data set, the larger the value of μ should be selected. According to our proposal if algorithm cannot reach a desired value after some iteration, μ will be set to 2 and rules that we mentioned above will be ignored (for special data sets). Because we use continues version of ICA, μ must be rounded to an integer, and then we must calculate ε by selected μ and relating to the dimensions of data set. Eq.(7) represents a method to calculate ε where rand is a random value between 0 and 1 with uniform distribution.

$$Eps = \sqrt{\mu} * rnd \tag{7}$$

According to the structure of ICA, introduced in section 3, algorithm continues to a specific number of iteration and calculates parameters for all countries at each iteration, then evaluates the results. For each iteration the best result will be compared to the existing optimum and finally we will have a global one. When the result of purity function which will be introduced in next section, reaches to 100%, algorithm will stop and represent results in diagrams and graphics.

4.1 Evaluation

Evaluation metrics are defined by mathematical and statistical functions. We know in clustering ideal metric must be considered by two aspects:

- **Cohesion:** intra-cluster patterns must have the most of similarity with each other.
- **Separation:** clusters must be considerably far from each other and the distance of cluster core must have the most possible distance.

Some of existing metrics to evaluate these aspects are Dunn metric [11], Calinski [12], Pakhira Bandyopadhyay Maulik (PBM) [13], Davies-Bouldin index (DB) [14] and Chou & Su (CS) [15] but they are not reliable for density-based clustering. In this section we will consider three functions to evaluate our proposed method. The first is purity function which has been used in many researches. This function evaluates obtained clusters with ε and μ, which calculated before by running OPTICS, to evaluate accuracy of our proposed method. The second is Means of Means (MoM) function to calculate the average distance of directly reachable points for every cluster. The third is standard deviation which is used to evaluate accuracy of parameters.

4.1.1 Purity function

After clustering with OPTICS data in clusters will be compared with class label and deviation will be calculated by purity function. The result will be considered as the result of goal function. Algorithm will improve the estimation of ε and μ until the value of purity function reaches 100% or algorithm violate maximum iteration. Calculation this method is presented in Eq.(8), where n is the number of objects in data set, k is the number of clusters which recognized by OPTICS and c_i is the number of objects which placed in cluster i by OPTICS and they are basically belong to cluster i.

$$Purity = \frac{100}{n}\sum_{i=1}^{k} \max(c_i) \,.\, c_i \in Index_{original} . Index_{ST-OPTICS} \tag{8}$$

4.1.2 Means of Means

To measure the distance of points in every cluster and the distance of clusters from each other we designed a function named Means of Means (MoM) which calculated

by the mean of directly reachable points. This function calculates the mean distance of directly reachable points for each and finally divides the result on the number of clusters to reach a normal value. The calculation is shown in Eq.(9) where n is the number of items within dataset, k is the number of clusters detected by OPTICS, c_i is the number of i^{th} cluster members which recognized by OPTICS, P_i are cores Q_i are points directly reachable by P_i and $|Q_i|$ is the number of directly reachable points by P_i.

$$Cohesion = \frac{\sum_{i=1}^{k}\sum_{j=1}^{c_i}(\frac{\|P_i.Q_i\|}{|Q_i|})}{k}. k * c = n \quad (9)$$

4.1.3 Standard deviation

Since clustering can be done correctly choosing different value for ε and μ, we must consider which value can present better clusters. To measure the accuracy of ε and μ, we will use standard deviation metric after clustering.

5. Experimental results

In this section, we shall use experimental results to show the clustering performance of ST-OPTICS. All experiments are implemented on a computer with Intel Pentium ® CPU 3.00 GHz, 8 GB of memory and 64Bit windows 8 operating system. All algorithms and data are implemented by matlab 8.3. To find optimized results the experiment was 10 times on every data sets and results are shown in tables 2.

5.1 Datasets

To evaluate the proposed method, we need data sets specialized for density based methods. In this paper we will use data sets from U.C.I. repository [16] and Mathworks site [17]. The numbers of 6 completely different data sets are considered. Since the density and shape of these data sets are different, they are suitable for our proposed method, ST-OPTICS, and can consider the efficiency of the method with high accuracy. Fig.6 shows the picture of these 6 data sets and Table.1 shows their statistics.

Table 1: statistics of data

Data set	***Class*** (#)	***Attribute***	***Size***
Aggregation	7	2	788(45,170,102,273,34,130,34)
Compound	6	2	399 (50,92,38,45,158,16)
Corners	4	2	1000 (250,250,250,250)
Crescent Full Moon	2	2	1000 (250,750)
Half Kernel	2	2	1000 (500,500)
R15	15	2	600 (40,40,40,40,40,40,40,40,40,40,40,40,40,40,40)

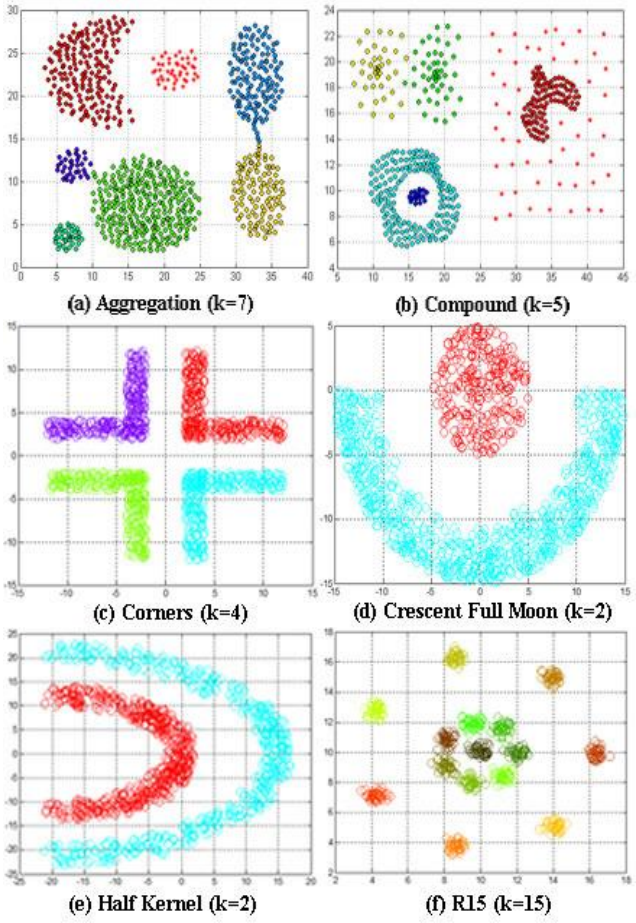

Fig 6: schema of data sets

5.2 The evaluation of proposed method

Table.2 shows the results of applying proposed method on values resulted from purity function, number of detected clusters, standard deviation and the value of ε and μ. Fig.7 to Fig.12 shows the pictures of Reachable-distance Dendrogram for all data sets.

Table 2: the evaluation of proposed method

Data set	***Purity (%)***	***Cluster (#)***	Standard deviation	μ	ε
Aggregation	99.952	7	0.040636	14	2.0218
Compound	87.3215	3	0.13737	2	1.541
Corners	100	4	0.17126	15	3.396
Crescent Full Moon	100	2	0.06769	7	1.2834
Half Kernel	100	2	0.07661	10	2.2798
R15	99.9341	15	0.021743	16	0.5741

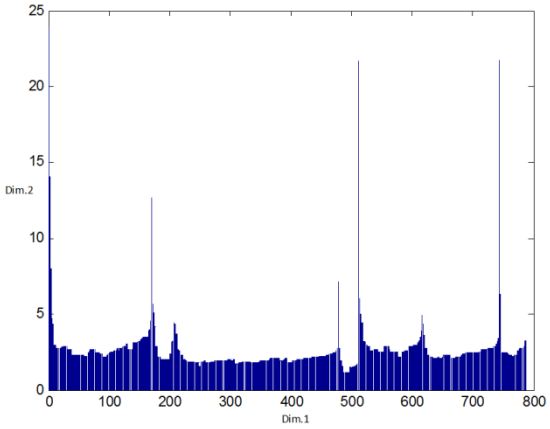

Fig 7: Reachable-distance Dendrogram for Aggregation k=7

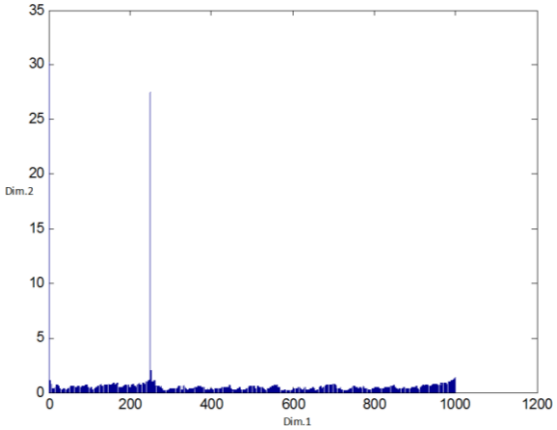

Fig 10: Reachable-distance Dendrogram for Crescent Full Moon k=2

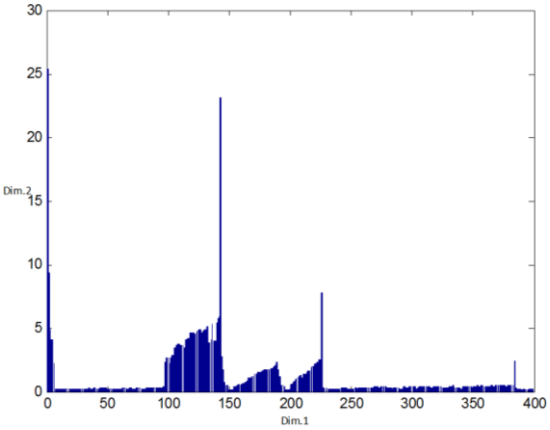

Fig 8: Reachable-distance Dendrogram for Compound k=5

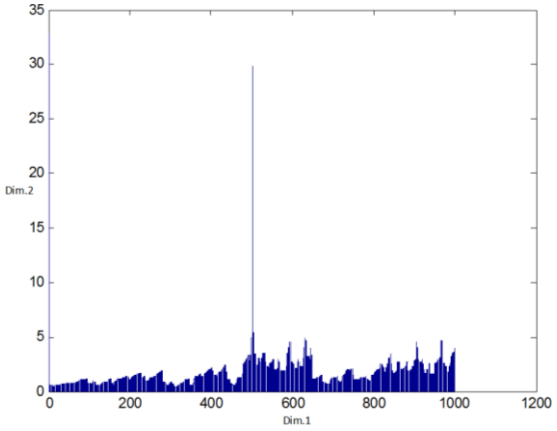

Fig 11: Reachable-distance Dendrogram for Half Kernel k=2

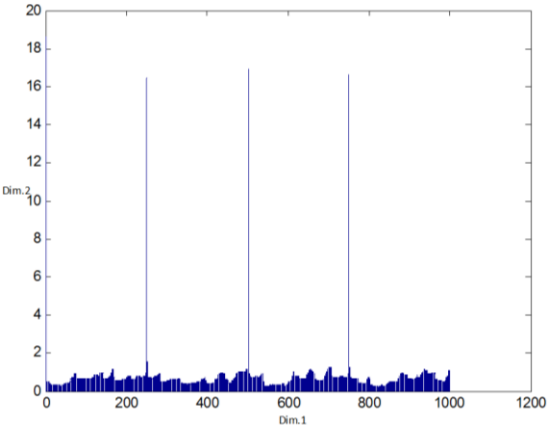

Fig 9: Reachable-distance Dendrogram for Corners k=4

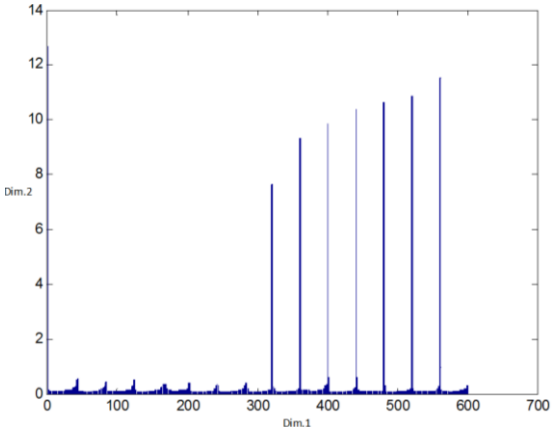

Fig 12: Reachable-distance Dendrogram for R15 k=15

6. Conclusions

This paper proposes a novel hybrid approach consisting Imperialist Competitive Algorithm and OPTICS clustering algorithm named as ST-OPTICS to choose quickly and automatically very well suited ε and μ parameters for OPTICS algorithm and improve its performance.

In order to evaluation of mentioned method we utilized different data sets specialized for classification and removed class label. Comparing our results with original one, proved that our method is produced a precise class label. According to the experimental results, our proposed method has the best results for any shapes of data sets and performed an optimized clustering.

Recently the use of meta-heuristic algorithms is growing in many fields of science but in data mining specially clustering, most researches focused on k-means clustering. This research represented the use of different meta-heuristic optimization algorithms and it seems that it is useful for many searching and optimization problems. In general, this method is very efficient for DBSCAN and SOM algorithms and problems with many parameters to set.

References

[1] C. J. Matheus, P. K. Chan, G. Piatetsky-Shapiro, Systems for knowledge discovery in databases, Knowledge and Data Engineering, IEEE Transactions on 5 (6) (1993) 903-913.
[2] N. Soni, A. Ganatra, Categorization of several clustering algorithms from different perspective: a review, International Journal of Advanced Research in Computer Science and Software Engineering 2 (8) (2012) 63-68.
[3] Ankerst, Mihael, et al. "OPTICS: ordering points to identify the clustering structure." ACM Sigmod Record. Vol. 28. No. 2. ACM, 1999.
[4] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise., in: Kdd, Vol. 96, 1996, pp. 226-231.
[5] M. C. Cowgill, R. J. Harvey, L. T. Watson, A genetic algorithm approach to cluster analysis, Computers & Mathematics with Applications 37 (7) (1999) 99-108.
[6] M. Mitchell, An introduction to genetic algorithms, MIT press, 1998.
[7] J. H. Holland, Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence., U Michigan Press, 1975.
[8] A. Colorni, M. Dorigo, V. Maniezzo, et al., Distributed optimization by ant colonies, in: Proceedings of the first European conference on artificial life, Vol. 142, Paris, France, 1991, pp. 134-142.
[9] S. Kirkpatrick, M. P. Vecchi, et al., Optimization by simmulated annealing, science 220 (4598) (1983) 671-680.
[10] E. Atashpaz-Gargari, C. Lucas, Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition, in: Evolutionary Computation, 2007. CEC 2007. IEEE Congress on, IEEE, 2007, pp. 4661-4667.
[11] J. C. Dunn, Well-separated clusters and optimal fuzzy partitions, Journal of cybernetics 4 (1) (1974) 95-104.
[12] T. Calinnski, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics-theory and Methods 3 (1) (1974) 1-27.
[13] M. K. Pakhira, S. Bandyopadhyay, U. Maulik, Validity index for crisp and fuzzy clusters, Pattern recognition 37 (3) (2004) 487-501.
[14] D. L. Davies, D. W. Bouldin, A cluster separation measure, Pattern Analysis and Machine Intelligence, IEEE Transactions on (2) (1979) 224-227.
[15] C.-H. Chou, M.-C. Su, E. Lai, A new cluster validity measure and its application to image compression, Pattern Analysis and Applications 7 (2) (2004) 205-220.
[16] archive.ics.usi.edu.
[17] www.mathworks.com.