IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 2, Ver. III (Mar – Apr. 2015), PP 56-62
www.iosrjournals.org

Particle Swarm Optimization based K-Prototype Clustering Algorithm

K. Arun Prabha[1], N. Karthi Keyani Visalakshi[2]
[1]*Assistant professor, Department of Computer Technology, Vellalar College for Women, Erode, Tamilnadu, INDIA.*
[2]*Associate Professor, Department of Computer Applications, Kongu Engineering College, Perundurai, Erode, Tamilnadu, INDIA.*

Abstract: *Clustering in data mining is a discovery process that groups a set of data so as to maximize the intra-cluster similarity and to minimize the inter-cluster similarity. The K-Means algorithm is best suited for clustering large numeric data sets when at possess only numeric values. The K-Modes extends to the K-Means when the domain is categorical. But in some applications, data objects are described by both numeric and categorical features. The K-Prototype algorithm is one of the most important algorithms for clustering this type of data. This algorithm produces locally optimal solution that dependent on the initial prototypes and order of object in the data. Particle Swarm Optimization is one of the simple optimization techniques, which can be effectively implemented to enhance the clustering results. But discrete or binary Particle Swarm Optimization mechanisms are useful for handle mixed data set. This leads to a better cost evaluation in the description space and subsequently enhanced processing of mixed data by the Particle Swarm Optimization. This paper proposes a new variant of binary Particle Swarm Optimization and K-Prototype algorithms to reach global optimal solution for clustering optimization problem. The proposed algorithm is implemented and evaluated on standard benchmark dataset taken from UCI machine learning repository. The comparative analysis proved that Particle Swarm based on K-Prototype algorithm provides better performance than the traditional K-modes and K-Prototype algorithms.*

Keywords*: Clustering, K-Means, K-Prototype Algorithm, Centroid, Particle Swarm Optimization.*

I. Introduction

Data clustering is a popular approach used to implement the partitioning operation and it provides an intelligent way of finding interesting groups when a problem becomes intractable for human analysis. It groups data objects based on the information found in the data that describes the objects and their relationships. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters[10]. Clustering has been studied in the field of machine learning and pattern recognition and it plays an important role in data mining applications such as scientific data exploration, information retrieval, opinion mining, and text mining. It also has a significant function in spatial database applications, web analysis, customer relationship management, social network data analysis, bio-medical analysis and many other related areas[3].

Clustering algorithm can be classified into Hierarchical clustering and Partitional clustering. Hierarchical clustering algorithm creates a hierarchical decomposition of data set, represented by a tree structure. Partitioning clustering constructs a partition of a given database of n data points into a predefined number of clusters. The partitional methods usually lead to better results because of its the nature of iterative and revised-type grouping method. The K-Means is one of the most widely used partitional clustering methods due to its simplicity, versatility, efficiency, empirical success and ease of implementation. This is evidenced by more than hundreds of publications over the last fifty five years that extend k-means in variety of ways[2].

The K-Means algorithm starts with K arbitrary centroids, typically chosen uniformly at random from the data objects[5]. Each data object is assigned to the nearest cluster centroid and then each centroid is recalculated as the mean of all data objects assigned to it. These two steps are repeated until a predefined termination criterion is met. The major handicap for K-Means is that it often limited to numerical data. Because, it typically uses Euclidean or squared Euclidean distance to measure the distortion between data objects and centroid and mean computation plays vital role in cluster identification. When mixed data are encountered, several researchers have applied different data transformation approaches to convert one type of attributes to the other, before executing K-Means algorithm. However, in some cases, these transformation approaches may result in loss of information, leading to undesired clustering results[8].

K-Modes[7] extends to the well known K-Means algorithm for clustering categorical data. This approach modifies the standard K-Means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centroids and

updating modes with the most frequent categorical values in each iteration of clustering process. Huang[9] proposed k-prototypes algorithm which is based on the k-means paradigm but removes the numeric data limitation whilst preserving its efficiency. This algorithm integrates the K-Means and K-Modes processes to cluster data with mixed numeric and categorical values. The random selection of starting centroids in these algorithms may lead to different clustering results and falling into local optima. Abundant algorithms have been developed to resolve this issue in K-Means by integrating excellent global optimization methods like Genetic algorithms (GA), ant colony optimization, Particle Swarm Optimization and etc.The Particle Swarm optimization (PSO)algorithms are randomized search and optimization techniques based on the concept of swarm. They are efficient, adaptive and robust search processes, performing multi-dimensional search in order to provide near optimal solutions of an evaluation (fitness) function in an optimization problem. In this paper, an attempt is made to integrate PSO with K- Prototype algorithm to reach global results while clustering categorical data.

Background

K-Prototype Clustering Algorithm

The K-Prototype algorithm integrates K-Means and K-Modes[19]. It is practically more useful for mixed-type objects. The dissimilarity between two mixed-type objects X and Y , which are described by attributes $A^r_{1,}$ $A^r_{2,\ldots\ldots}$ $A^r_{p,}$ $A^r_{p+1,\ldots\ldots}$ $A^c_{m,}$, can be measured by

$$d_2 = \sum_{j=1}^{p}(x_j - z)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_{j,} z) \qquad (1)$$

where the first term is the squared Euclidean distance measure on the numeric attributes and the second term is the simple matching dissimilarity measure on the categorical attributes. Selection of a □ □ value is guided by the average standard deviation (□ □ □ of numeric attributes[20]. A suitable □ □ to balance the similarity measure is between 0.5 and 0.7. A suitable □ □ lies between 1 and 2. Therefore,a suitable □ □ lies between 1/3□ □ and 2/3□ □ for these data sets[18]. Based on this approach, work to be done to find the value of □ . The influence of γ in the clustering process is given in Table-2.

According to Huang[6] the cost function for mixed-type objects is as follows:

$$J = \sum_{l=1}^{k} (P_l^r + P_l^c) \qquad (2)$$

Where $$p_l^r = \sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{i,j} - z)^2 \qquad (3)$$

$$p_l^c = \gamma \sum_{i=1}^{n} w_{i,l} \sum_{j=p+1}^{m} (x_{i,j} - z) \qquad (4)$$

The above (4) has written as

$$J = \sum_{l=1}^{k} (\sum_{i=1}^{n} w_{i,l} \sum_{j=1}^{p} (x_{i,j} - z)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_{i,j}, z)) \qquad (5)$$

Since both P_l^r and P_l^c are nonnegative, minimizing J is equivalent to minimizing P_l^r and P_l^c for 1<=l<=k.

ALGORITHM : K-Prototype Clustering Algorithm
Step 1: Select k initial prototypes for k clusters from the data set X.
Step 2: Allocate each data object in X to the cluster whose prototype is the nearest to it according to (1).Update the prototype of the cluster after each allocation.
Step 3: After all data objects have been allocated to a cluster, Retest the similarity of data objects against the current prototypes. If a data object is found that its nearest prototype belongs to another cluster rather than its current one, reallocate the data object to that cluster and update the prototypes of both clusters.
Step 4: Repeat Step 3 until no data object has changed clusters after a full cycle test of X.

Particle Swarm Optimization:

PSO is an efficient and effective global optimization algorithm, which can be used to solve multimodal, non-convex and noncontiguous problems[12]. A Particle is individual object and when a number of particles are grouped, it is termed as swarm. PSO is associated with velocity. Particles fly through the search space with

velocities dynamically adjusted velocities as per their historical behaviors. The particles therefore have the tendency to fly towards the better and better search area all over the course of the process of search. The particles try to achieve to global minimum by using global and local best information[3]. PSO is working based on the intelligence and the search can be carried out by the speed of the particle. The PSO algorithm operates iteration by iteration and solution produced in each iteration is compared with self-local best and global best of swarm.

PSO algorithm consists of following steps:

ALGORITHM : PSO Algorithm
Step 1 : Initialize each particle with random position and Velocity
Step 2 : Evaluate the fitness of each particle
Step 3 : Update p_{best} and G_{best} of each particle
Step 4 : Update velocity and position of each particle using (6) and (7) respectively

$$V_p(t+1) = w * V_p(t) + c_1 * r_1 * (P_{best} - X_p(t)) + c_2 * r_1 * (G_{best} - X_p(t)) \quad (6)$$

$$X_p(t+1) = X_p(t) + V_p(t+1) \quad (7)$$

Step 5 : Terminate till the condition met.

The inertia weight w is calculated for each iteration using (8).

$$W = (W_{\max} - W_{\min} / \max no.\ of\ iterations) * t \quad (8)$$

The main advantage of PSO is that it has less parameter to adjust and fast convergence, when it is compared with many global optimization algorithms like Genetic algorithms (GA), Simulated Annealing (SA) and other global optimization algorithms.

Discrete or Binary Particle Swarm Optimization

PSO is designed for continuous function optimization problems[17]. It is not used for discrete function optimization problems. To overcome this problem discrete or binary PSO is proposed. The major difference between binary PSO and ordinary PSO is that the velocities and positions of the particles are defined in terms of the changes of probabilities and the particles are formed by integers in {0,1}. Therefore a particle flies in a search space restricted to zero or one. The speed of the particle must be considered to the interval [0,1]. A logistic sigmoid transformation function $S(v_i(t+1))$ is shown in the following equation,

$$S(v_i(t+1)) = \frac{1}{1+e\wedge(v_i(t+1))} \quad (9)$$

The new position of the particle is obtained by using the following equation $X_p(t+1) = 1$ if $r_3 < S(v_i(t+1))$ otherwise 0.Where r_3 is the uniform random number in the range [0,1].

Related Research

This section reviews various algorithms proposed for mixed numeric and categorical data and recently published Particle Swarm Optimization based K-Means algorithms.

Zhexue Huang[6] proposed K-Modes to introduce new dissimilarity measures and to deal with categorical objects, which replace the means of clusters with modes and activate the use a frequency based method to update modes in the clustering process to minimize the clustering cost function. Zhexue Huang and Michael K. Ng[20] formulated a fuzzy K-Mode approach to the K-Means paradigm to cluster large categorical data sets efficiently. Michael K. Ng, Mark Junjie Liy Joshua, Zhexue Huang and Zengyou He[18] has derived the updating formula of the K-Modes clustering algorithm with the new dissimilarity measure for the convergence of the algorithm under the optimization framework.

Zhexue Huang[7] proposed a method to dynamically update the K-Prototypes in order to maximize the intra cluster similarity of objects. An improved multi-level clustering algorithm based on k-prototype proposed by LI Shi-jin, ZHU Yue-long, LIU Jing[13]**.** The low purity problem was occurred when k-prototype algorithm was working to process complex data sets. To tackle this issue, the new algorithm was proposed. In order to improve the quality of clustering, re-clustering was performed on those clusters with low-purity through automatic selection of attributes. Extension to the K-Prototypes algorithms, hard and fuzzy K is proposed by Wei-Dong Zhao, Wei-Hui Dai, and Chun-Bin Tang[21]. It focuses on effects of attribute values with different frequencies on clustering accuracy to propose new update method for centroids. A fuzzy K-Prototype clustering algorithm for mixed numeric and categorical data are proposed by Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han, Zhe Wang[11]. In this paper, mean and fuzzy centroid are combined to represent the prototype of a cluster, and employed in a new measure based on co-occurrence of values, to evaluate the dissimilarity between data objects and prototypes of clusters. This measure also takes into account the significance of different attributes towards the clustering process. An algorithm for clustering mixed data is formulated.

An improved k-prototype clustering algorithm for mixed numeric and categorical data proposed by Jinchao Ji, Tian Bai, Chunguang Zhou, Chao Ma, Zhe Wang[10]. In this paper, the concept of the distribution centroid for representing the prototype of categorical attributes in a cluster was introduced. Then combine both mean with distribution centroid to represent the prototype of the cluster with mixed attributes, and thus propose a new measure to calculate the dissimilarity between data objects and prototypes of clusters. This measure takes into account the significance of different attributes towards the clustering process for mixed datasets. Izhar Ahmad[9] compared the performance of K-Means and K-Prototype Algorithm. In this research a detail discussion of the K-Means and K-Prototype to recommend efficient algorithm for outlier detection and other issues relating to the database clustering. The verification and validation of the system is based on the simulation.

R. Madhuri, M. Ramakrishna Murty, J. V. R. Murthy, P. V. G. D. Prasad Reddy, Suresh C. Satapathy[14] proposed two algorithms namely K-Modes and K-Prototype algorithms for clustering categorical data sets. And also reduce the cost functions.

Particle Swarm Optimization based K-Means clustering approach for security assessment in power systems was proposed by S.Kalyani, K.S.Swarup[12]. This paper demonstrates how the traditional K-Means clustering algorithm can be profitably modified to be used as a classifier algorithm. The proposed algorithm combines the Particle Swarm Optimization (PSO) with the traditional K-Means algorithm to satisfy the requirements of a classifier. Omar S. Soliman, Doaa A. Saleh, and Samaa Rashwan, T. Huang et al[19] proposed bio inspired fuzzy K-modes clustering algorithm. It integrates concepts of FK-Modes algorithm to handle the uncertainty phenomena and FPSO to reach global optimal solution of clustering optimization problem. K. Arun Prabha, N. Karthikeyani Visalakshi[3] proposed an effective partitional clustering algorithm which is developed by integrating the merits of Particle Swarm Optimization and normalization with traditional K-Means clustering algorithms. Improved global-best particle swarm optimization algorithm with mixed-attribute data classification capability proposed by Nabila Nouaoria and Mounir Boukadoum[17]. In this algorithm, a new particle-position update mechanism is proposed to handle mixed data. This interpretation mechanism uses the frequencies of non numerical attributes. This enhanced algorithm gives better cost function and processing of mixed attribute data.

Proposed Algorithm

PSO based K-Prototype Clustering Algorithm

K-Prototype Clustering is an effective algorithm for clustering mixed type data sets. The dependency of the algorithm on the initialization of the centers is a major problem and its usually gets stuck in local optima. To solve this issue, PSO and K-Prototype algorithms are combined. The proposed algorithm does not depend on the initial clusters. It can be avoided by being trapped in local optimal solutions. In this method, the process is initialized with a group of random population N. A population is called a Swarm.

The PSO based K-Prototype algorithms consists of the following steps :

ALGORITHM : PSO based K-Prototype Clustering Algorithm

Input : Data of n objects with d features, PSO Parameters c_1, p, r_1, $r2$, W_{max}, W_{min}, the value of γ and the value of K

Output : K clusters

Procedure

Step 1 : Initialize a population of particle with small random positions, x_p and velocities, v_p of the pth particle on problem space of K x D dimensions.

Step 2 : Initialize the PSO parameters c_1, c_2, r_1, r_2, p, w_{max} and w_{min}.

Step 3 : Repeat the step 4 to step 11

Step 4 : Start the procedure and set the iterative count t=1.

Step 5 : Run the following steps in the K-Prototype algorithm ,

For every object in the population

i. Calculate Squared Euclidean distance measure for numeric data

ii. Calculate the simple matching dissimilarity measure on the categorical attributes.

iii. Assign each data object to nearest cluster center.

Step 6: After grouping the data objects based on the minimum distance, Calculate the cost function using (5) for every object

Step 7 : Compute p_{best} based on the cost function

Step 8 : After updating $p_{best,}$ choose the best value among the particles in p_{best} and assign to G_{best} ,i.e., If $p_{best} < G_{best}$, then $G_{best} = p_{best}$

Step 9 : Modify the velocity using the equations (6) and update the new position of each particle using the equations (7) and (9) respectively.

Step 10 : Compute t = t + 1

Step 11 : Check the convergence criteria, which may be a good fitness value or a maximum number of iterations.

A Swarm consists of X particles moving around a D-dimensional search space. Given a data set X = $\{x_1, x_2, x_3 \ldots\ldots, x_N\}$, where x_i is a data pattern in a D dimensional feature space, each particle is of dimension K x D, K being the number of clusters for partitioning the data set X. The position of the i_{th} object is represented by $x_i = (x_{i1}, x_{i2}, x_{i3,\ldots} x_{iD,})$ and its velocity is represented as $v_i = (v_{i1}, v_{i2}, v_{i3,\ldots} v_{iD})$ where i is the index of the object and D is the dimensionality of the search space. PSO maintains a population of particles, each one characterized by a position vector in the search space and a velocity vector which determines its motion. The velocity is calculated based on i)particle's current direction, ii) each particle is attracted to the best position it has achieved so far and iii) each particle is attracted to the best particle in population. Each object declares its best velocity by p_{best} and the best value in group as G_{best}. Here K-Prototype clustering algorithm is executed to find the optimum value. The Squared Euclidean distance is calculated for every numeric object in the datasets. For categorical attribute Huang's simple dissimilarity measure is applied. The evaluation of the previous p_{best} value is compared with the current p_{best} value in terms of cost function. If the current position is better than p_{best} it is contended as G_{best} or else the previous of the p_{best} can be retained as G_{best}. The position and velocity of the i_{th} object are updated by p_{besti} and G_{best} in the each generation. After finding the two best values, the particle updates its velocity and position with the equations (6) and (7) respectively. The value of inertia weight w is calculated using the equation(8). The inertia weight w controls the impact of the previous velocity. The particles cannot fly continuously through a discrete-valued space. So here the discrete values of the particles are converted into the continuous values based on discrete or binary PSO. The new position update mechanism is implemented by the equation (9). For every object in the dataset K-Prototype cost function is calculated using the Equation(5) and the p_{best} is found. This leads to better cost function evaluation in the description space. This enhanced procedure proposed to handle mixed data attributes with the PSO. Here the typical value of c1 and c_2 is taken as 2.0, r_1 and r_2 is a random number generated between 0 and 1. A linearly decreasing inertia weight (w) was implemented by starting at $w_{max} = 0.9$ and $w_{min} = 0.2$. This helps to expand the search space in the beginning so that the particles can explore new areas, implying a global search. The □ value can be formulated based on the average standard deviation (□ □ of numeric attributes. Therefore, a suitable □ □ lies between 1/3□ □ and 2/3□ □ for these data sets[15]. In this paper the value of □ □ is□ evaluated and it stands between 0.5 and 0.7. The proposed method enhances the convergence speed of PSO and aids in tracing the initial centriod K-Prototype clustering algorithm. The swarm based algorithms aids to analyze the global optimal solutions.

II. Results And Discussions

The experiment analysis is performed with Hepatitis, Post operative patient, Australian Credit Approval, German Credit Data and Stat log Heart benchmark data sets available in the UCI machine learning repository[15]. The details of the data sets are given in the following Table-1.

The performance of K-Modes, K-Prototype and PSO based K-Prototype algorithm is measured in terms of four external validity measures namely Rand Index, Jaccard Index , F-Measure and Entropy. The external validity measures test the quality of clusters by comparing the results of clustering with the 'ground truth' (true class labels). All these four measures have a value between 0 and 1. In case of Rand Index, Jaccard Index and F-Measure, the value 1 indicates that the data clusters are exactly same and so increase in the values of these measures proves the better performance.

Table-1: Details of datasets

S. No.	Dataset	No. of Instances	No. of Attributes	No. of Classes
1.	Hepatitis	155	19	2
2.	Post operative patient	90	8	3
3.	Australian Credit Approval	690	14	2
4.	German Credit Data	1000	20	2
5.	Statlog Heart	270	13	2

The results of PSO based K-Prototype clustering algorithm, in comparison with the results of K-Modes algorithm, K-Prototype algorithm in terms of Rand Index, Jaccard Index , Entropy and F-Measure are shown in Table-3, Table-4, Table-5 and Table-6 respectively.By means of analysis the details of the data sets and corresponding lambda values for the five benchmark datasets are shown in the following Table-2.

According to Rand Index, the performance of PSO-KP clustering yields consistent and improved results than K-Modes and K-Prototype Algorithm in almost all datasets. From the Table-3 it is observed that PSO-KP algorithm yields consistent and better results for Stat log Heart ,Hepatitis, German Credit Data than Post operative patient, and Australian Credit Approval data sets.

Table-2: Details of data sets and lamda value

Dataset	Hepatitis	Post operative patient	Australian Credit Approval	German Credit Data	Statlog Heart
Lambda vaue	0.0533	0.1055	0.0680	0.0995	0.0845

Table-3: Comparative analysis based on Rand index

S. No	Dataset	K-Modes	K-Prototype	PSO based K-Prototype
1.	Hepatitis	0.6234	0.6171	0.6723
2.	Post operative patient	0.4815	0.4822	0.4998
3.	Australian Credit Approval	0.6565	0.6853	0.6971
4.	German Credit Data	0.5000	0.5144	0.5312
5.	Statlog Heart	0.5988	0.7029	0.7429

From the Table-4, based to Jaccard Index, the performance of PSO-KP algorithm yields consistent and better results for data sets. K-Modes and K-Prototype Algorithm in almost all datasets.

Table-4:Comparative analysis based on Jaccard Index

S. No.	Dataset	K-Modes	K-Prototype	PSO based K-Prototype
1.	Hepatitis	0.5225	0.6099	0.6723
2.	Post operative patient	0.3598	0.3870	0.3950
3.	Australian Credit	0.5019	0.5240	0.5326
4.	German Credit Data	0.3839	0.3959	0.4141
5.	Statlog Heart	0.4424	0.5505	0.5705

In case of F-Measure, the value 1 indicates that the data clusters are exactly same and so the increase in the values of these measures proves the better performance. Based on this, the results of PSO-KP is appreciable than K-Modes and K-Prototype algorithm for all datasets represented in Table 5.

Table-5: Comparative analysis based on F-measure

S. No	Dataset	K-Modes	K-Prototype	PSO based K-Prototype
1.	Hepatitis	0.7266	0.7217	0.7521
2.	Post operative patient	0.5486	0.5615	0.5752
3.	Australian Credit	0.7829	0.8039	0.8229
4.	German Credit Data	0.5921	0.6041	0.6261
5.	Statlog Heart	0.7258	0.8197	0.8387

The decrease in the values of Entropy measure proves the better performance. Based on that the performance of PSO-KP based on Entropy is highly significant than K-Modes and K-Prototype for all dataset except Australian Credit Approval represented in Table-6.

Table-6 : Comparative analysis based on Entropy

S. No	Dataset	K-Modes	K-Prototype	PSO based K-Prototype
1.	Hepatitis	0.4060	0.3625	0.3421
2.	Post operative patient	0.6662	0.6635	0.6231
3.	Australian Credit	0.5267	0.4896	0.5725
4.	German Credit Data	0.6012	0.6090	0.5961
5.	Statlog Heart	0.5550	0.4735	0.4635

Based on the comparative analysis, it is concluded that that PSO-KP algorithm proves better performance for all experimented mixed numeric and categorical datasets. It shows the superiority of the proposed algorithm to produce the optimal number of clusters.

III. Conclusion

This paper proposed PSO based K-Prototype Clustering algorithm by incorporating the benefit of PSO with the existing K-Prototype algorithm, to reach the global optimum cluster solution. The proposed algorithm has been tested on the five benchmark data sets which include both numeric and categorical attributes. It is proved that the performance of the proposed algorithm is superior to the performance of conventional K-Modes and K-Prototype clustering algorithms. In future, appropriate optimization algorithm will be applied for tuning of parameter to produce superior quality clusters. The global cluster results can further be improved by setting alternate values for the parameters of PSO.

References

[1]. Amir Ahmad and Lipika Dey, Algorithm for Fuzzy Clustering of Mixed Data with Numeric and Categorical Attributes, Proceedings of ICDCIT, 561-572 **(2005**)

[2]. Anil K. Jain, Data Clustering: 50 Years Beyond K-Means, Pattern Recognition Letters, **(2009)**

[3]. Arun Prabha K. and Karthikeyani Visalakshi N., Improved Particle Swarm Optimization based K-Means Clustering, International Conference on Intelligent Computing Applications, **(2014)**

[4]. Han J. and Kamber M., Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, San Francisco**, (2006)**

[5]. He Z., Xu X., Deng S., Scalable Algorithms for Clustering Mixed Type Attributes in Large datasets, International Journal of Intelligent Systems, **20**, 1077-1089 **(2005)**

[6]. Huang Z., Extensions to the K-Means algorithm for clustering large data sets with categorical values, DataMining and Knowledge Discovery , **2(3),** 283-304 **(1998)**

[7]. Huang Z., Clustering large data sets with mixed numeric and categorical Values, Proceedings of the FirstAsia Confernce on Knowledge Discovery and Data Mining, 21-34 **(1997)**

[8]. Huang Z., A fast clustering algorithm to cluster very large categorical data sets in data mining, proceedings of the SIGMOD Workshop on Research Issues on Data Mining and Know ledge Discovery, Tucson, Arizona, USA, 1-8 **(1997a)**

[9]. Izhar Ahmad, K-Mean and K-Prototype Algorithms Performance Analysis, American Research Institute for Policy Development, **2(1),** 95-109 (2014).

[10]. Jinchao Ji., Tian Bai., Chunguang Zhou., Chao Ma., Zhe Wang., An Improved K-Prototypes clustering algorithm for mixed numeric and categorical data, Image Feature Detection and Description, **20**, 590-596 **(2013)**

[11]. Jinchao Ji., Wei Pang., Chunguang Zhou., Xiao Han., Zhe Wang., A fuzzy K-Prototype clustering algorithm for mixed numeric and categorical data, Knowledge-Based Systems, **30**, 129-135 **(2012)**

[12]. Kalyani S. and Swarup K.S., Particle swarm optimization based K-Means clustering approach for security assessment in power systems, Expert systems with applications, **38**, 10839-10846 **(2011)**

[13]. LI Shi-jin, ZHU Yue-long, LIU Jing, An improved multi-level clustering algorithm based on k-prototype, Journal of Software, **(2005)**

[14]. Madhuri R., Ramakrishna Murty M., Murthy J.V.R , Prasad Reddy P.V.G.D., Suresh Satapathy C., Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms, Advances in Intelligent Systems and Computing, **249**, 137-144 **(2014)**

[15]. Merz C.J., Murphy P.M., UCI repository of machine learning data bases, Irvine, Universityof California, http://www.ics.uci.eedu/~mlearn/, (**199**8**)**

[16]. Ming-Yi Shih, Jar-Wen Jheng, Lien-Fu Lai, A Two-Step Method for Clustering Mixed Categorical and Numeric Data, Tamkang Journal of Science and Engineering, **13(1),** 11-19 **(2010)**

[17]. Nabila Nouaouria , Mounir Boukadoum, Improved global-best particle swarm optimization algorithm with mixed-attribute data classification capability , Applied Soft Computing, **21**, 554–567 **(2014)**

[18]. Ng, Li M.J., Huang J.Z., He Z., On the impact of dissimilarity measure in K-Modes clustering algorithm , IEEE Transactions on Pattern Analysis and Machine Intelligence , **29(3)** , 503–507 **(2007)**

[19]. Omar S. Soliman., Doaa A. Saleh., and Samaa Rashwan., A Bio Inspired Fuzzy K-Modes Clustring Algorithm , ICONIP, Part III, LNCS 7665, 663–669 **(2012)**

[20]. Sotirios P. Chatzis, A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functions, Expert Systems with Applications, **38**, 8684–8689 **(2011)**

[21]. Wei-Dong Zhao, Wei-Hui Dai, Chun-Bin Tang, K-Centrs Algorithm for Clustering Mixed type data, PAKDD, 1140-1147 **(2007)**

[22]. Zhexue Huang and Michael K. Ng, A Fuzzy K-Modes Algorithm for Clustering Categorical Data, IEEE Transactions on Fuzzy Systems, **7(4)** , 446-452

[23]. Zhi Zheng , Maoguo Gong , Jingjing Ma,; Licheng Jiao ,Unsupervised evolutionary clustering algorithm for mixed type data , Evolutionary Computation (CEC), 1-8 **(2010**)