Developments and Trends in Intelligent Technologies and Smart Systems

Vijayan Sugumaran
Oakland University, USA

A volume in the Advances in Computational Intelligence and Robotics (ACIR) Book Series

IGI Global
DISSEMINATOR OF KNOWLEDGE

Published in the United States of America by
IGI Global
Engineering Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue
Hershey PA, USA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com

Library of Congress Cataloging-in-Publication Data

Names: Sugumaran, Vijayan, 1960- editor.
Title: Developments and trends in intelligent technologies and smart systems / Vijayan Sugumaran, editor.
Description: Hershey, PA : Engineering Science Reference, [2018] | Includes bibliographical references.
Identifiers: LCCN 2017020847| ISBN 9781522536864 (hardcover) | ISBN 9781522536871 (ebook)
Subjects: LCSH: Systems engineering. | Self-organizing systems.
Classification: LCC TA168 .D496 2018 | DDC 620/.0042--dc23 LC record available at https://lccn.loc.gov/2017020847

This book is published in the IGI Global book series Advances in Computational Intelligence and Robotics (ACIR) (ISSN: 2327-0411; eISSN: 2327-042X)

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Chapter 10
Clustering Mixed Datasets Using K-Prototype Algorithm Based on Crow-Search Optimization

Lakshmi K.
Kongu Engineering College, India

Karthikeyani Visalakshi N.
NKR Government Arts College for Women, India

Shanthi S.
Kongu Engineering College, India

Parvathavarthini S.
Kongu Engineering College, India

ABSTRACT

Data mining techniques are useful to discover the interesting knowledge from the large amount of data objects. Clustering is one of the data mining techniques for knowledge discovery and it is the unsupervised learning method and it analyses the data objects without knowing class labels. The k-prototype is the most widely-used partitional clustering algorithm for clustering the data objects with mixed numeric and categorical type of data. This algorithm provides the local optimum solution due to its selection of initial prototypes randomly. Recently, there are number of optimization algorithms are introduced to obtain the global optimum solution. The Crow Search algorithm is one the recently developed population based meta-heuristic optimization algorithm. This algorithm is based on the intelligent behavior of the crows. In this paper, k-prototype clustering algorithm is integrated with the Crow Search optimization algorithm to produce the global optimum solution.

DOI: 10.4018/978-1-5225-3686-4.ch010

INTRODUCTION

Knowledge Discovery in Databases (KDD) is an automatic, exploratory analysis and modelling of large data repositories. It is the organized as the process of identifying valid, novel, useful, and understandable patterns from large and complex data sets. Data Mining is the heart of the KDD process, involving the large number of algorithms that explore the data, develop the model and discover previously unknown patterns.

Data clustering is the process of grouping the heterogeneous data objects into homogeneous clusters such that data objects within the cluster are similar with each other and dissimilar between the other clusters.

Clustering is used in variety of fields like data mining and knowledge discovery, market research, machine learning, biology, pattern recognition, weather prediction, etc. An early specific example of the use of cluster analysis in market research is given in (Green, Frank & Robinson, 1967). A large number of cities were used as test markets and the cluster analysis was used to classify the cities into a small number of groups on the basis of variables includes city size, newspaper circulation and per capita income. It shows that cities within a group is very similar to each other, choosing one city from each group was used for selecting the test markets.

Another example is, Littmann (2000) applies cluster analysis to the daily occurrences of several surface pressures for weather in the Mediterranean basin, and finds the groups that explain rainfall variance in the core Mediterranean regions. Liu and George (2005) use fuzzy k-means clustering to account for the spatiotemporal nature of weather data in the South-Central USA. Kerr and Churchill (2001) investigate the problem of clustering tools applied to gene expression data.

There are number of clustering algorithms are available for grouping the instances of the same type. The clustering algorithms are categorized into Partitional clustering algorithms, Hierarchical clustering algorithms, Density-Based clustering algorithms and Grid-Based clustering algorithms. Partitional clustering algorithms form the clusters by partition the data objects into groups. Hierarchical clustering algorithms form the clusters by the hierarchical decomposition of data objects.

The partitional clustering algorithms include k-means, k-modes, k-medoids and k-medians. The hierarchical clustering algorithms can be classified as single linkage and complete linkage, agglomerative algorithms. Density based clustering algorithms can be listed as DBSCAN, DENCLUE, OPTICS. The grid based clustering algorithms include GRIDCLUS, BANG and STING.

The k-means algorithm handles the large amount of data objects but it handles numeric type data objects. Huang introduced the two extensions of the k-means clustering algorithm. First extension is the k-modes clustering algorithm (Huang, 1997a) and second extension is the k-prototype clustering algorithm (Huang, 1997b). The k-modes algorithm efficiently handles the large amount of categorical data objects. The k-prototype algorithm efficiently handles the large amount of data objects with numeric and categorical types of data objects. This algorithm is the integration of k-means and k-modes clustering algorithms. For the mixed numeric and categorical datasets, the Euclidean distance is calculated for numeric data and the matching similarity measure is calculated for categorical data.

The k-prototype clustering algorithm selects the initial prototypes randomly from the data objects and it leads to the local optimum solution. To overcome this problem, optimization algorithm is integrated with k-prototype clustering algorithm.

Recently, there are number of optimization algorithms are introduced to obtain the global optimum solution. Some of the nature-inspired metaheuristic optimization algorithms are Genetic Algorithm (GA)

(Holland, 1975; Goldberg, 1989), Ant Colony Optimization (ACO) (Dorigo, 1992), Simulated Annealing (SA) (Brooks & Morgan, 1995), Particle Swarm Optimization (PSO) (Eberhart & Kennedy, 1995), Tabu Search (TS) (Glover & Laguna, 1997), Cat Swarm Optimization (CSO) (Chu, Tsai & Pan, 2006), Artificial Bee Colony (ABC) (Basturk & Karaboga, 2006), Cuckoo Search (CS) (Yang & Deb, 2009, 2010), Gravitational Search (GS) (Rashedi, Nezamabadi-Pour & Saryazdi, 2009), Firefly Algorithm (FA) (Yang, 2010), Bat Algorithm (BA) (Yang, 2010), Wolf Search Algorithm (WSA) (Tang, Fong, Yang & Deb, 2012), Krill Herd (KH) (Gandomi & Alavi, 2012).

Crow Search Algorithm (CSA) (Askarzadeh, 2016) is one of the metaheuristic population based optimization algorithms and it was introduced by Askarzadeh in 2016. This algorithm simulates the intelligent behavior of the crows. Crows are considered as one of the world's most intelligent birds. This algorithm is based on finding the hidden storage position of excess food of crows. Finding food source is hidden by another crow is not easy task because if a crow finds any one following it, the crows tries to fool the crow by moving to another position. This algorithm is very simple and easy to understand. Each optimization algorithm has controlling parameters to achieve the performance of the algorithms. Also, the number of controlling parameters for CSA algorithm is two namely awareness probability and flight length.

The reason behind this work is k-prototype algorithm produces the local optimum solution. Also, Huang (1997b) suggested the global optimization for the k-prototype algorithm. To overcome the k-prototype local optimum problem, this paper Crow Search optimization algorithm combined with the k-prototype clustering algorithm.

The organization of this paper is as follows: Section 2 describes the related researches in the literature. Section 3 describes the k-prototype clustering algorithm. Section 4 describes the Crow Search Algorithm. Section 5 describes the proposed algorithm. The experimental analysis is discussed in Section 6. Conclusion and future works are provided in Section 7.

RELATED WORK

Ant Colony Optimization approach for clustering problem is given in (Shelokar, Jayaraman & Kulkarni, 2004). Simulated Annealing algorithm approach for clustering algorithms is proposed in (Selim & Alsultan, 1991). Particle Swarm Optimization approach for clustering problem is given in (Chen & Ye, 2004). Tabu Search algorithm approach for clustering problem is proposed in (Al-Sultan, 1995). Artificial Bee Colony Optimization approach for clustering algorithms is given in (Zhang, Ouyang & Ning, 2010; Karaboga & Ozturk, 2011). Cat Swarm Optimization approach for clustering problem is given in (Santosa, & Ningrum, 2009).

Genetic Algorithm combined with k-means was proposed in (Krishna & Murty, 1999). Hybrid clustering algorithm based on k-means and ant colony algorithm was proposed in (Lu & Hu, 2013). Cluster analysis with k-means and Simulated Annealing was introduced in (Sun, Xu, Liang, Xie, & Yu, 1994). Particle Swarm Optimization based k-means clustering algorithm was proposed in (Van der Merwe & Engelbrecht, 2003; Ahmadyfard & Modares, 2008). Tabu Search based k-means was developed in (Liu, Liu, Wang & Chen, 2005). Artificial Bee Colony based k-means algorithm was proposed in (Armano & Farmani, 2014). Gravitational Search algorithm, combined with k-means was introduced in (Hatamlou, Abdullah & Nezamabadi-Pour, 2012). Firefly Algorithm is combined with k-means was proposed in (Hassanzadeh & Meybodi, 2012). Bat Algorithm is combined with k-means was proposed in (Komarasamy & Wahi, 2012). Wolf Search Algorithm, Cuckoo Search, Bat Algorithm, Firefly Algorithm and Ant Colony Optimization algorithms are integrated with k-means in introduced in (Tang, Fong, Yang & Deb, 2012).

Tabu search algorithm is combined with k-modes is introduced in (Ng & Wong, 2002). Genetic Algorithm is combined with k-modes is developed in (Gan, Yang & Wu, 2005). It finds the global optimum solution for the given categorical dataset and the crossover operator is replaced with k-modes operator. Fuzzy based k-modes algorithm is proposed in (Huang, & Ng, 1999). In hard clustering, each data object is assigned to single cluster. In fuzzy clustering, each object belongs to more than one cluster and the membership degree value is varying from one cluster to another. The fuzzy k-modes algorithm integrated with Genetic Algorithm for categorical data was proposed in (Gan, Wu & Yang, 2009). It treated the fuzzy k-modes algorithm as an optimization problem and Genetic Algorithm is used to obtain the global optimum solution.

Swarm-based k-modes algorithm is introduced in (Izakian, Abraham & Sná, 2009). A novel approach for combining Particle Swarm Optimization with k-modes is proposed in (Mei & Xiang-Jun, 2012). First, the categorical data are mapped to natural numbers, find the similarity between the data objects and initial centroids and finally update the mode by using the frequency based method.

The Particle Swarm Optimization algorithm integrated with k-modes clustering algorithm and this hybridized algorithm is applied to retrieve the three dimensional objects was proposed in (Zhao & Lu, 2013). Artificial Bee Colony based k-modes is developed in (Ji, Pang, Zheng, Wang & Ma, 2015). In this paper, one-step k-modes clustering algorithm procedure is executed and then integrate this procedure with the artificial bee colony approach.

Yin & Tan (2005) proposed the new way of clustering mixed numeric and categorical type of data objects. In this paper, proposed the improved k-prototype clustering algorithm. For clustering, first step is use the *CF**-tree to pre-cluster datasets. After the dense regions are stored in leaf nodes, then each dense region as a single point and use an improved k-prototype to cluster such dense regions.

Ahmad and Dey (2007) proposed the new cost function for clustering mixed numeric and categorical attributes. It provides the cost for both numeric and categorical attributes. It is computed from each attribute from the given data objects. But Huang provides the cost only for categorical attributes. Also apply a new distance method between two categorical attribute values. In this, the new distance is computed from the overall distribution of values in a single class and the overall distribution of values in the dataset.

The evolutionary k-prototypes (EKP) algorithm by (Zheng, Gong, Ma, Jiao & Wu, 2010) integrates the evolutionary framework with k-prototype algorithm. In this paper, proposed the Evolutionary based k-prototype algorithm for mixed numeric and categorical datasets. The cross over operator and mutation operator is applied separately for each kind of data. Also apply the simulated binary crossover operator for numerical and single point crossover for categorical data object in the dataset. Also apply the polynomial mutation for numerical data objects and uniform mutation for categorical data objects in the dataset. The tournament selection with elitism strategy is used for selecting the individuals for each generation.

Chatzis (2011) introduce an extension of the GG algorithm to allow for the effective handling of data with mixed numeric and categorical attributes. Traditionally, fuzzy clustering of such data is conducted by means of the fuzzy *k*-prototypes algorithm, which merely consists in the execution of the original FCM algorithm using a different dissimilarity functional, suitable for attributes with mixed numeric and categorical attributes.

Pham, Suarez-Alvarez, and Prostov (2011) developed the new clustering algorithm called RANKPRO that is combines the honey bee optimization algorithm with k-prototype clustering algorithm. The honey bee algorithm uses the random search method instead of using genetic algorithm operators like crossover and mutation. Also apply the normalization procedure to balance the sum of numeric and categorical attributes and avoid either type of attribute.

Ji, Pang, Zhou, Han and Wang (2012) proposed the fuzzy based k-prototype algorithm for clustering mixed numeric and categorical datasets. In this paper, fuzzy c-mean type clustering algorithm for mixed numeric and categorical attributes is presented. In this algorithm, combination of mean and fuzzy centroids to represent prototype for a cluster and apply the new mew measure based on co-occurrence of values to assess the dissimilarity between the data objects and prototypes of clusters.

Ji, Bai, Zhou, Ma & Wang (2013) proposed the improved k-prototype clustering algorithm for mixed numeric and categorical attributes is proposed. In this algorithm, introduce the distribution centroids to represent the prototypes of cluster with mixed attributes and propose the new measure to assess the dissimilarity between the data objects and prototypes of clusters. The new measure is based on the Huang strategy of evaluate the significance of the attributes in the dataset.

Wu Sen, Chen Hong, and Feng Xiaodong (2013) proposed a new dissimilarity measure for incomplete data set with mixed numeric and categorical attributes and a new approach to select k objects as the initial prototypes based on the nearest neighbors. The improved k-prototypes algorithm cluster incomplete data without need to impute the missing values, randomness in choosing initial prototypes.

Madhuri, Murty, Murthy, Reddy and Satapathy (2014) implemented algorithms which extend the k-means algorithm to categorical domains by using modified k-modes algorithm and domains with mixed categorical and numerical values by using k-prototypes algorithm k-prototypes algorithm which is implemented by integrating the Incremental k-means and the Modified k-modes partition clustering algorithms.

Ji et al., (2015) propose a novel cluster center initialization method for the k-prototypes algorithms to address this issue. In the proposed method, the centrality of data objects is introduced based on the concept of neighbor-set, and then both the centrality and distance are exploited together to determine initial cluster centers.

Prabha and Visalakshi (2015) proposed the particle swarm optimization based k-prototype algorithm. In this paper, binary particle swarm optimization is integrated with the k-prototype clustering algorithm to obtain the global optimum solutions.

Lakshmi, Visalakshi and Shanthi (2017) proposed the cuckoo search based k-prototype algorithm. In this work, cuckoo search optimization algorithm is integrated with the k-prototype clustering algorithm to obtain the global optimum solutions.

In (Arun & Kumar, 2017) applied the Artificial Bee Colony (ABC) optimization algorithm for online analytical query processing in data warehouse. The authors apply the ABC algorithm for OLAP to minimize the query response time. Also proposed the Artificial Bee Colony (ABC) based view selection algorithm.

Krishnamoorthy, Sadasivam, Rajalakshmi, Kowsalyaa and& Dhivya (2017) proposed Particle Swarm Optimization based system is to hide a group of interesting patterns which contains sensitive knowledge. This system also reduces the side effects like number of modifications.

Naser and Alshattnawi (2014) proposed the new way to group the social networks based on Artificial Bee Colony optimization algorithm, which is a swarm based meta-heuristic optimization algorithm. This approach aims to maximize the modularity, which is a measure that represents the quality of network partitioning.

K-PROTOTYPE ALGORITHM

The k-prototype algorithm (Huang, 1997b) is the partition based clustering algorithm that clustering the data objects with both the numeric and categorical and also efficiently handles the large amount of data objects.

Let X = {x_{11}, x_{12},...,x_{nm}}be the data object with n number of instances with m attributes. Let k is the number clusters given by the user. The objective of k-prototype clustering algorithm is to divide the n number of data objects into k number of clusters and minimize the cost function defined in the following equation (1):

$$E\left(U,Q\right)=\sum_{l=1}^{k}\sum_{i=1}^{n}u_{il}dis\left(x_i,Q_l\right) \tag{1}$$

u_{il} the element of the partition matrix U_{nxk}; Q_l is the prototype of cluster l; x_i is the data object. The dis(x_i,Q_l) is calculated using the following equation (2):

$$dis\left(x_i,Q_l\right)=\sum_{j=1}^{p}\left(x_{ij}^{r}-q_{lj}^{r}\right)+\alpha\sum_{j=p+1}^{m}\delta\left(x_{ij}^{c}-q_{lj}^{c}\right) \tag{2}$$

$\sum_{j=1}^{p}\left(x_{ij}^{r}-q_{lj}^{r}\right)$ is the Euclidean distance between the data objects and the prototype of cluster for numeric attributes. The Euclidean distance is calculated using the equation (3):

$$d_{num}(x,y)=\sqrt{\sum_{i=1}^{m}\left(x_i-y_i\right)^2} \tag{3}$$

$\sum_{j=p+1}^{m}\delta\left(x_{ij}^{c}-q_{lj}^{c}\right)$ is the matching dissimilarity measure between the data objects and the prototype of cluster for categorical attributes. The α specifies the weight for categorical attributes. The matching dissimilarity is evaluated using the equation (4):

$$d_{cat}(x,y)=\begin{cases}0, & x_i=y_i\\ 1, & x_i\neq y_i\end{cases} \tag{4}$$

The k-prototype clustering algorithm is described as follows:

Step 1: Randomly select k initial prototypes as the initial cluster centres from the dataset X.
Step 2: For each data object in X, calculate the distance between the data object and the initial centroids using the equation (1)

Step 3: Assign the data objects to cluster whose data object have the minimum distance.
Step 4: After the initial assignment of data objects to clusters, update the initial prototype based on the newly assigned data objects using the equation (1).
Step 5: Repeat the step 4 until no changes in the clustership of data objects.

CROW SEARCH ALGORITHM

The Crow Search Algorithm (CSA) (Askarzadeh, 2016) mimics the intelligent and foraging behaviour of the crows. The crow follows the other crows to steal the food hidden by that crows. The principles of crow search algorithm are (i) They live in the form of groups (ii) remember the position of food hiding locations (iii) follow the each other for stealing food (iv) protect their food source. In CSA, diversification and intensification are controlled by the Awareness Probability (AP).

The number of crows i.e flock size is P with D attributes and the position of the crow is i at time in iteration in the search space is specified as $X_{i,iter}$, i = 1, 2, …, N; iter = 1, 2, …, itmax, itmax is the maximum number iterations. Each crow has a memory m to remember the position of the hiding place. At each iteration, the position of food hidden place for crow i is specified by $m_{i,iter}$ and it shows the best position obtained so far.

The CSA is described as follows:

```
Step 1: Initialize number of flocks P, maximum number of iterations itmax,
flight length fl, awareness probability AP.
Step 2: Initialize the position of each crow randomly with PxD dimensional
search space. Initialize the memory of the crows with the initial position of
crows.
Step 3: Evaluate the position of the crows.
       a. While iter < maxiter
          i. for all crows
             1. Randomly choose any one of the crows to follow (for example μ)
             2. If crow μ does not know that crow ν is following it, new
                position of μ is obtained using the equation (5):
```

$$x^{i,iter+1} = x^{i,it} + r_i \times fl^{i,it} \times \left(m^{j,iter} - x^{i,it}\right) \tag{5}$$

```
             3. If crow μ does know that crow ν is following it, new position of
                μ is obtained by the randomly using the following equation (6):
```

$$x^{i,iter+1} = \text{a random position} \tag{6}$$

```
             4. The equations (5) and (6) is combined in the following
                equation (7):
```

$$\begin{cases} x^{i,it} + r_i \times fl^{i,it} \times \left(m^{j,iter} - x^{i,it} \right) & r_j \geq AP^{j,iter} \\ \text{a random position} & \text{otherwise} \end{cases} \tag{7}$$

5. Check the feasibility of the new position. If the new position of crow is feasible, its position is updated, otherwise the crow stays in the current position.
6. Evaluate the new position of the crows
7. Update the memory of the crows by using the equation (8):

$$\begin{cases} x^{i,it+1} & f\left(x^{i,it+1}\right) is\, better\, than\, f\left(m^{i,it}\right) \\ m^{i,iter} & otherwise \end{cases} \tag{8}$$

b. End of while

PROPOSED ALGORITHM

The k-prototype clustering algorithm is the combination of k-means and k-modes clustering algorithm. Both the k-means and k-modes clustering algorithms are efficiently handling large amount of numeric and categorical data respectively. The k-prototype algorithm also efficiently handling large amount of mixed numeric and categorical datasets. The main drawback of this algorithm is producing local optimum solutions. To obtain the global optimum solutions, k-prototype is combined with global optimization algorithms. The Crow Search algorithm is the population based metaheuristic optimization algorithm and it mimics the intelligent behaviour of the crows. In this proposed work, Crow Search Algorithm combined with k-prototypes algorithm to obtain the global optimum solution.

Algorithm Steps

Step 1: Input the datasets X with the N number data objects with D number of attributes, Number of clusters K, Flock size P, Maximum number of iterations maxiter, flight length fl, and awareness probability AP.
Step 2: Initialize the position of crows for P by generating the matrix with the random numbers with the size of P rows with KxD columns. The maximum range of random numbers is the total number of instances in the data objects.
Step 3: Encode the random numbers with the data objects. Each row specifies the K cluster center for clustering algorithm.
Step 4: Initialize the memory of the crows with the values of the initial position of the crows because initially crows hidden their foods in their initial positions.
Step 5: Evaluate the fitness of initial position of crows by using the equation (1).

```
Step 6: Initialize the fitness of memory of the crows with the fitness posi-
tion of the crows.
Step 7: Update the position of crows:
       a. while iteration<=maxiter
          i. for all crows
             1. Choose any one of the crows to follow randomly (for example μ).
             2. If crow μ does not know that crow ν is following it, new
                position of μ is obtained using the equation (5).
             3. If crow μ does know that crow ν is following it, new position
                of μ is obtained by the randomly using the equation (6).
             4. Check the feasibility of the new position. If the new position
                of crow is feasible, its position is updated, otherwise the
                crow stays in the current position.
       b. End of while
Step 8: Evaluate the fitness of new position of crows by using the equation (1).
Step 9: Update the memory of the crows by using the equation (8).
Step 10: Finally, the best position G_best is obtained.
Step 11: Run the k-prototype algorithm with G_best as the prototype for clusters.
Step 12: Calculate the Euclidean distance for numeric data and matching simi-
larity for categorical data from each data to G_best obtained from CSA.
Step 13: Repeat Step 12 until convergence criteria is met.
```

EXPERIMENTAL RESULTS

The algorithms are implemented using Matlab R2015a on Intel i5 2.30 GHz with 4GB RAM. The k-prototypes, PSOk-prototypes and CSAk-prototypes are executed 20 distinct runs. The algorithm specific parameters are specified in Table 1. The values for the Particle Swarm Optimization algorithm are suggested in (Van den Bergh, 2001). The values for the Crow Search algorithm are suggested in (Askarzadeh, 2016).

Table 1. Algorithm specific parameters

Criteria	k-prototype	PSOk-prototype	CSAk-prototype
Iterations	20	100	100
Particles	N/A	15	15
Parameters	$\alpha = 0.5$	$w = 0.72$ $c1 = 1.49$ $c2 = 1.49$	$fl = 2$ $AP = 0.1$

Datasets

The proposed CSAk-prototype clustering algorithm is tested with the benchmark mixed datasets such as Bupa, Credit Approval, Heart, Hepatitis, Post-Operative Patient and Zoo. These datasets are are obtained from the UCI machine learning repository (Asuncion & Newman, 2007). The details of these datasets are described in the Table 2. In this work, with the help of standard metrics such as FMeasure, Accuracy and Rand Index to assess the quality of the clustering results.

Measures

For all measures, use the four terms namely, TP, TN, FP and FN. TP means True Positive, it is the count of actual and predicted values are same. TN means True Negative and the actual and predicted values are different. A FP means False Positive, decision means that values with different features are assigned to the same cluster. A FN means False Negative, decision means that the values with similar traits to different clusters. N is the total number of objects.

The FMeasure (Van Rijsbergen, 1979) is an external index. It is the harmonic mean of the precision and recall coefficients. If the precision is high and recall value is low, this results in a low FMeasure. If both precision and recall are low, a low FMeasure is obtained. On the other hand, if both are high, a high FMeasure value is obtained. FMeasure can be computed using the formula (9):

$$FMeasure = 2 \times \frac{\text{Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}} \tag{9}$$

Precision is calculated as the number of correct positive predictions divided by the total number of positive predictions. The best precision is 1, whereas the worst is 0. Precision is calculated as true positive divided by the sum of false positive and true positive. It is calculated using the equation (10):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{10}$$

Table 2. Details of datasets

Dataset	No. of Instances	No. of Attributes	No. of Numeric Attributes	No. of Categorical Attributes	No. of Classes
Bupa	345	6	5	1	2
Credit Approval	690	15	6	9	2
Heart	270	13	6	7	2
Hepatitis	155	19	6	13	2
Post Operative Patient	90	8	1	7	3
Zoo	101	16	1	15	7

Recall is calculated as the number of correct positive predictions divided by the total number of positives. The best sensitivity is 1.0, whereas the worst is 0.0. It is calculated using the equation (11):

$$\text{Recall} = \frac{TP}{N} \tag{11}$$

Accuracy is calculated as the number of all correct predictions divided by the total number of the data objects. In case of accuracy, the value 1 indicates data object is clustered exactly same. Highest value of this measure indicates better performance. It is calculated using the equation (12):

$$Accuracy = \frac{TP + TN}{N} \tag{12}$$

Rand Index (Rand, 1971) is a measure of the similarity between true labels and predicted labels. It is calculated using the equation (13):

$$RandIndex = \frac{TP + TN}{TP + FP + TN + FN} \tag{13}$$

The Rand index has the value lies between 0 and 1, 0 indicating that the two data clusters do not agree on any pair of points and 1 indicating that the data clusters are exactly the same.

Results

The following Table 3 shows the best, worst, average and standard deviation of objective function values for the benchmark datasets.

For bupa, hepatitis and zoo datasets, the CSAk-prototype algorithm outperforms compared to k-prototype and PSOk-prototype algorithms. That is, best, worst, average and standard deviation values are better than k-prototype and PSOk-prototype algorithms.

For credit approval and Post-Operative Patient datasets, the CSAk-prototype algorithm outperforms compared to k-prototype and PSOk-prototype algorithms. The best, worst and average values are better than k-prototype and PSOk-prototype algorithms. But the standard deviation of k-prototype is better than PSOk-prototype and CSAk-prototype algorithms.

For heart dataset, the CSAk-prototype algorithm outperforms compared to k-prototype and PSOk-prototype algorithms. That is worst, average values and standard deviation are better than k-prototype and PSOk-prototype algorithms. But the best value of k-prototype is better than PSOk-prototype and CSAk-prototype algorithms.

The following Table 4 shows the FMeasure, Accuracy, and Rand Index scores for the benchmark datasets.

For Bupa dataset, k-prototype algorithm gives the best FMeasure compare with PSOk-prototypes and CSAk-prototypes algorithms. CSAk-prototype algorithm gives the best accuracy and rand index values compare with k-prototypes and PSOk-prototypes algorithms. For Credit Approval, CSAk-prototype

Table 3. Comparison of objective function values obtained from three algorithms

Dataset	Criteria	k-prototype	PSOk-prototype	CSAk-prototype
Bupa	Best	10485.9368	10463.1151	**9974.8872**
	Worst	14955.1969	11472.9706	**10944.1463**
	Average	10806.3294	10549.9688	**10518.6544**
	Std	1150.9185	265.6908	**184.7416**
Credit Approval	Best	544174.2443	538637.2660	**534710.9724**
	Worst	666640.7309	677683.5431	**657656.5320**
	Average	595637.6748	594218.9867	**592486.4524**
	Std	**23750.1241**	29164.7700	24935.0278
Heart	Best	**11025.7623**	11033.0958	11060.3928
	Worst	17158.3642	12732.6091	**11742.8632**
	Average	11488.5261	11176.5074	**11107. 2340**
	Std	1568.6460	431.1003	**175.8933**
Hepatitis	Best	9225.7405	9057.2432	**9050.6731**
	Worst	13914.3494	9994.0377	**9990.3276**
	Average	9887.5603	9708.6395	**9687.2750**
	Std	1119.6490	204.6099	**190.6814**
Post Operative Patient	Best	194.1308	118.0000	**116.0000**
	Worst	214.8000	212.8529	**200.0833**
	Average	197.1877	192.3803	**189.0477**
	Std	**6.5806**	27.4801	20.2992
Zoo	Best	131.5000	112.5000	**102.0000**
	Worst	252.2500	249.1818	**196.8548**
	Average	224.1747	185.2364	**170.7366**
	Std	29.0243	33.6960	**21.0479**

Table 4. Comparison of FMeasure, accuracy and RandIndex of three algorithms

Dataset	FMeasure			Accuracy			Rand Index		
	k-prototype	PSOk-prototype	CSAk-prototype	k-prototype	PSOk-prototype	CSAk-prototype	k-prototype	PSOk-prototype	CSAk-prototype
Bupa	**0.5568**	0.5551	0.5561	55.16	55.71	**55.87**	0.5040	0.5051	**0.5055**
Credit Approval	0.6358	0.6384	**0.6392**	66.84	67.03	**67.08**	0.5560	0.5573	**0.5576**
Heart	0.6018	0.6024	**0.6045**	60.09	60.12	**60.33**	0.5187	0.5188	**0.5196**
Hepatitis	**0.6465**	0.6460	0.6433	61.16	61.06	**60.74**	**0.5219**	0.5214	0.5214
Post Operative Patient	0.5129	**0.5445**	0.5223	45.52	**49.25**	46.32	0.4844	**0.4937**	0.4862
Zoo	0.6535	0.6809	**0.6899**	59.90	62.82	**64.51**	0.8221	0.8416	**0.8459**

algorithm gives the best FMeasure, accuracy and rand index compare with k-prototypes and CSAk-prototypes algorithms.

For Heart, CSAk-prototype algorithm gives the best FMeasure, accuracy and rand index values when compared with k-prototypes and CSAk-prototypes algorithms. For Hepatitis, k-prototype algorithm gives the best FMeasure compare with PSOk-prototypes and CSAk-prototypes algorithms. CSAk-prototype algorithm gives the best accuracy and rand index compare with k-prototypes and PSOk-prototypes algorithms. For Post Operative Patient, PSOk-prototype algorithm gives the best FMeasure, accuracy and rand index compare with k-prototypes and CSAk-prototypes algorithms. For Zoo, CSAk-prototype algorithm gives the best FMeasure, accuracy and rand index compare with values when compared with k-prototypes and PSOk-prototypes algorithms.

The Figures 1 to 3 show the overall performance of Accuracy, FMeasure and RandIndex of k-prototypes, PSOk-prototypes and CSAk-prototypes algorithms.

Comparison of CSA With PSO

All optimization algorithms have individual controlling parameters. But the number of parameters is varying from one to another algorithm. Parameter setting is the time-consuming task and lagging in setting the proper values for algorithms. In PSO, requires four parameters like maximum velocity, inertia weight, social learning factor and individual learning factor. In CSA, requires two parameters like flight length and awareness probability.

In PSO, have the complexity like need to initialize and check the boundaries of velocity. If the velocity is reached below minimum and it is set to the minimum velocity. If the velocity is reached beyond the upper maximum and it is set to the maximum velocity. In CSA, need to check the upper and lower bounds of newly obtained position of the crow. If the position is greater than lower bound and less than upper bound, it is set to the new position of the crow.

Figure 1. Comparison of accuracy values of three algorithms

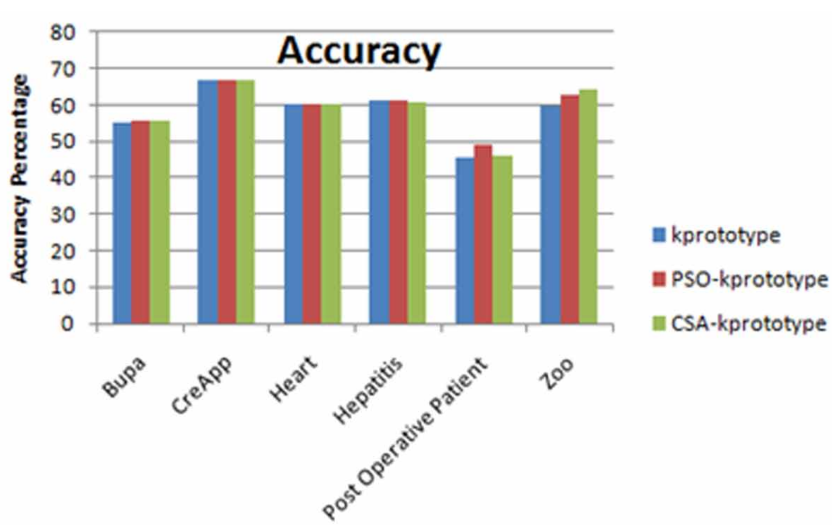


Figure 2. Comparison of FMeasure values of three algorithms

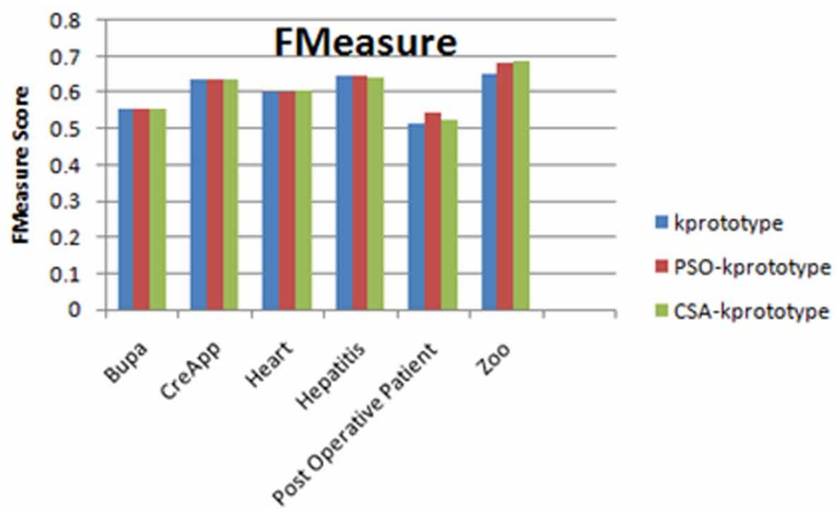
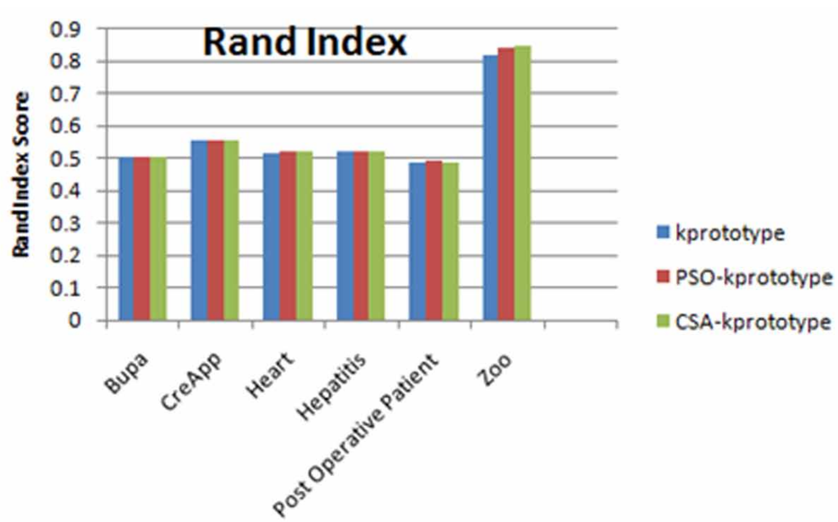


Figure 3. Comparison of RandIndex values of three algorithms



Both the PSO and CSA have the memories to maintain the good solutions. In PSO, each particle attracted towards the best positions maintained in its memory. In CSA, at each iteration, each crow selects randomly one of the flock crows to move towards its hiding place. The best positions found are directly used to find the better position.

CONCLUSION

This work is motivated by the problem of clustering large mixed datasets because most of the datasets are mixed numeric and categorical. Mixed datasets are ubiquitous in real world database. However, few efficient algorithms are available for clustering mixed numeric and categorical data objects. The k-prototype clustering algorithm is easy to implement and efficiently handling large numeric and categorical datasets. In this paper, incorporate the k-prototype clustering algorithm with Crow Search Optimization algorithm to obtain the global optimum solution. The efficiency of the proposed algorithm is experimented with six benchmark datasets and the results are compared with k-prototype and Particle Swarm Optimization with k-prototype algorithms. The experimental results show that the Crow Search algorithm with k-prototype is outperforms for Credit approval, heart and Zoo datasets than k-prototype and Particle Swarm Optimization with k-prototype algorithms. It also shows that the PSO with k-prototype is outperforms for Post Operative Patient than k-prototype and Crow Search algorithm with k-prototype algorithms. In future, extend this work with measure the clustering results with internal validity measures.

REFERENCES

Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data & Knowledge Engineering*, *63*(2), 503–527. doi:10.1016/j.datak.2007.03.016

Ahmadyfard, A., & Modares, H. (2008). Combining PSO and k-means to enhance data clustering. In *Proceedings of the International Symposium on Telecommunications IST '08* (pp. 688-691). IEEE. doi:10.1109/ISTEL.2008.4651388

Al-Sultan, K. S. (1995). A tabu search approach to the clustering problem. *Pattern Recognition*, *28*(9), 1443–1451. doi:10.1016/0031-3203(95)00022-R

Armano, G., & Farmani, M. R. (2014). Clustering analysis with combination of artificial bee colony algorithm and k-means technique. *International Journal of Computer Theory and Engineering*, *6*(2), 141–145. doi:10.7763/IJCTE.2014.V6.852

Arun, B., & Kumar, T. V. (2017). Materialized View Selection using Artificial Bee Colony Optimization. *International Journal of Intelligent Information Technologies*, *13*(1), 26–49. doi:10.4018/IJIIT.2017010102

Askarzadeh, A. (2016). A novel metaheuristic method for solving constrained engineering optimization problems: Crow search algorithm. *Computers & Structures*, *169*, 1–12. doi:10.1016/j.compstruc.2016.03.001

Asuncion, A., & Newman, D. (2007). UCI machine learning repository.

Basturk, B., & Karaboga, D. (2006). An artificial bee colony (abc) algorithm for numeric function optimization. In *Proceedings of the IEEE Swarm Intelligence Symposium 2006*, Indianapolis, Indiana, USA.

Brooks, S. P., & Morgan, B. J. (1995). Optimization using simulated annealing. *The Statistician*, *44*(2), 241–257. doi:10.2307/2348448

Chatzis, S. P. (2011). A fuzzy c-means-type algorithm for clustering of data with mixed numeric and categorical attributes employing a probabilistic dissimilarity functional. *Expert Systems with Applications*, *38*(7), 8684–8689. doi:10.1016/j.eswa.2011.01.074

Chen, C. Y., & Ye, F. (2004). Particle swarm optimization algorithm and its application to clustering analysis. In *Proceedings of the 2004 IEEE International Conference on Networking, Sensing and Control* (Vol. 2, pp. 789-794). IEEE.

Chu, S. C., Tsai, P. W., & Pan, J. S. (2006). Cat swarm optimization. In *Proceedings of the Pacific Rim International Conference on Artificial Intelligence* (pp. 854-858). Springer Berlin Heidelberg.

Dorigo, M. (1992). Optimization, learning and natural algorithms [Ph.D thesis]. Politecnico di Milano, Italy.

Eberhart, R. C., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In *Proceedings of the sixth international symposium on micro machine and human science* (Vol. 1, pp. 39-43). doi:10.1109/MHS.1995.494215

Gan, G., Wu, J., & Yang, Z. (2009). A genetic fuzzy k-Modes algorithm for clustering categorical data. *Expert Systems with Applications*, *36*(2), 1615–1620. doi:10.1016/j.eswa.2007.11.045

Gan, G., Yang, Z., & Wu, J. (2005). A genetic k-modes algorithm for clustering categorical data. In *Proceedings of the International Conference on Advanced Data Mining and Applications* (pp. 195-202). Springer Berlin Heidelberg. doi:10.1007/11527503_23

Gandomi, A. H., & Alavi, A. H. (2012). Krill herd: A new bio-inspired optimization algorithm. *Communications in Nonlinear Science and Numerical Simulation*, *17*(12), 4831–4845. doi:10.1016/j.cnsns.2012.05.010

Glover, F., & Laguna, M. (1997). *Tabu search*. Boston: Kluwer Academic Publishers. doi:10.1007/978-1-4615-6089-0

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley.

Green, P. E., Frank, R. E., & Robinson, P. J. (1967). Cluster analysis in test market selection. *Management Science*, *13*(8), 387–400. doi:10.1287/mnsc.13.8.B387

Hassanzadeh, T., & Meybodi, M. R. (2012). A new hybrid approach for data clustering using firefly algorithm and K-means. In *Proceedings of the 2012 16th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP)* (pp. 007-011). IEEE.

Hatamlou, A., Abdullah, S., & Nezamabadi-Pour, H. (2012). A combined approach for clustering based on K-means and gravitational search algorithms. *Swarm and Evolutionary Computation*, *6*, 47–52. doi:10.1016/j.swevo.2012.02.003

Holland, J. (1975). *Adaption in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press.

Huang, Z. (1997a). A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining. In DMKD (p. 0).

Huang, Z. (1997b). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining, (PAKDD)* (pp. 21-34).

Huang, Z., & Ng, M. K. (1999). A fuzzy k-modes algorithm for clustering categorical data. *IEEE Transactions on Fuzzy Systems*, *7*(4), 446–452. doi:10.1109/91.784206

Izakian, H., Abraham, A., & Sná, V. (2009). Clustering categorical data using a swarm-based method. In *Proceedings of the World Congress on Nature & Biologically Inspired Computing NaBIC '09* (pp. 1720-1724). IEEE. doi:10.1109/NABIC.2009.5393623

Ji, J., Bai, T., Zhou, C., Ma, C., & Wang, Z. (2013). An improved k-prototypes clustering algorithm for mixed numeric and categorical data. *Neurocomputing*, *120*, 590–596. doi:10.1016/j.neucom.2013.04.011

Ji, J., Pang, W., Zheng, Y., Wang, Z., & Ma, Z. (2015). A novel artificial bee colony based clustering algorithm for categorical data. *PLoS ONE*, *10*(5), e0127125. doi:10.1371/journal.pone.0127125 PMID:25993469

Ji, J., Pang, W., Zheng, Y., Wang, Z., Ma, Z., & Zhang, L. (2015). A novel cluster center initialization method for the k-prototypes algorithms using centrality and distance. *Applied Mathematics & Information Sciences*, *9*(6), 2933.

Ji, J., Pang, W., Zhou, C., Han, X., & Wang, Z. (2012). A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data. *Knowledge-Based Systems*, *30*, 129–135. doi:10.1016/j.knosys.2012.01.006

Karaboga, D., & Ozturk, C. (2011). A novel clustering approach: Artificial Bee Colony (ABC) algorithm. *Applied Soft Computing*, *11*(1), 652–657. doi:10.1016/j.asoc.2009.12.025

Kerr, M. K., & Churchill, G. A. (2001). Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments. *Proceedings of the National Academy of the USA*, *98*(16), 8961–8965. doi:10.1073/pnas.161273698 PMID:11470909

Komarasamy, G., & Wahi, A. (2012). An optimized K-means clustering technique using bat algorithm. *European Journal of Scientific Research*, *84*(2), 26–273.

Krishna, K., & Murty, M. N. (1999). Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, *29*(3), 433–439. doi:10.1109/3477.764879 PMID:18252317

Krishnamoorthy, S., Sadasivam, G. S., Rajalakshmi, M., Kowsalyaa, K., & Dhivya, M. (2017). Privacy Preserving Fuzzy Association Rule Mining in Data Clusters Using Particle Swarm Optimization. *International Journal of Intelligent Information Technologies*, *13*(2), 1–20. doi:10.4018/IJIIT.2017040101

Lakshmi, K., Visalakshi, N. K., & Shanthi, S. (2017). Cuckoo Search based K-Prototype Clustering Algorithm. *Asian Journal of Research in Social Sciences and Humanities*, *7*(2), 300–309. doi:10.5958/2249-7315.2017.00092.2

Littmann, T. (2000). An empirical classification of weather types in the Mediterranean Basin and their interrelation with rainfall. *Theoretical and Applied Climatology*, *66*(3-4), 161–171. doi:10.1007/s007040070022

Liu, S., & George, R. (2005). *Mining Weather Data using Fuzzy Cluster Analysis*. Berlin: Springer. doi:10.1007/3-540-26886-3_5

Liu, Y., Liu, Y., Wang, L., & Chen, K. (2005). A hybrid tabu search based clustering algorithm. In *Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems* (pp. 186-192). Springer Berlin Heidelberg. doi:10.1007/11552451_25

Lu, J., & Hu, R. (2013). A new hybrid clustering algorithm based on K-means and ant colony algorithm. In *Proceedings of the 2nd International Conference on Computer Science and Electronics Engineering*. doi:10.2991/iccsee.2013.430

Madhuri, R., Murty, M. R., Murthy, J. V. R., Reddy, P. P., & Satapathy, S. C. (2014). Cluster Analysis on Different Data Sets Using K-Modes and K-Prototype Algorithms. In *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II* (pp. 137-144). Springer International Publishing. doi:10.1007/978-3-319-03095-1_15

Mei, L., & Xiang-Jun, Z. (2012). A Novel PSO k-Modes Algorithm for Clustering Categorical Data. In Computer, Informatics, Cybernetics and Applications (pp. 1395-1402). Springer Netherlands. doi:10.1007/978-94-007-1839-5_150

Naser, A. M. A., & Alshattnawi, S. (2014). An Artificial Bee Colony (ABC) Algorithm for Efficient Partitioning of Social Networks. *International Journal of Intelligent Information Technologies*, *10*(4), 24–39. doi:10.4018/ijiit.2014100102

Ng, M. K., & Wong, J. C. (2002). Clustering categorical data sets using tabu search techniques. *Pattern Recognition*, *35*(12), 2783–2790. doi:10.1016/S0031-3203(02)00021-3

Pham, D. T., Suarez-Alvarez, M. M., & Prostov, Y. I. (2011). Random search with k-prototypes algorithm for clustering mixed datasets. In *Proceedings of the Royal Society of London A: mathematical, physical and engineering sciences* (Vol. 467, No. 2132, pp. 2387-2403). The Royal Society. doi:10.1098/rspa.2010.0594

Prabha, K. A., & Visalakshi, N. K. (2015). Particle Swarm Optimization based K-Prototype Clustering Algorithm. *Journal of Computer Engineering*, *1*(17), 56–62.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850. doi:10.1080/01621459.1971.10482356

Rashedi, E., Nezamabadi-Pour, H., & Saryazdi, S. (2009). GSA: A gravitational search algorithm. *Information Sciences*, *179*(13), 2232–2248. doi:10.1016/j.ins.2009.03.004

Santosa, B., & Ningrum, M. K. (2009). Cat swarm optimization for clustering. In *Proceedings of the International Conference of Soft Computing and Pattern Recognition SOCPAR '09* (pp. 54-59). IEEE. doi:10.1109/SoCPaR.2009.23

Selim, S. Z., & Alsultan, K. (1991). A simulated annealing (SA) algorithm for the clustering problem. *Pattern Recognition*, *24*(10), 1003–1008. doi:10.1016/0031-3203(91)90097-O

Shelokar, P. S., Jayaraman, V. K., & Kulkarni, B. D. (2004). An ant colony approach for clustering. *Analytica Chimica Acta*, *509*(2), 187–195. doi:10.1016/j.aca.2003.12.032

Sun, L. X., Xu, F., Liang, Y. Z., Xie, Y. L., & Yu, R. Q. (1994). Cluster analysis by the K-means algorithm and simulated annealing. *Chemometrics and Intelligent Laboratory Systems*, *25*(1), 51–60. doi:10.1016/0169-7439(94)00049-2

Tang, R., Fong, S., Yang, X. S., & Deb, S. (2012). Integrating nature-inspired optimization algorithms to K-means clustering. In *Proceedings of the 2012 Seventh International Conference on Digital Information Management (ICDIM)* (pp. 116-123). IEEE.

Tang, R., Fong, S., Yang, X. S., & Deb, S. (2012). Wolf search algorithm with ephemeral memory. In *Proceedings of the 2012 Seventh International Conference on Digital Information Management (ICDIM)* (pp. 165-172). IEEE. doi:10.1109/ICDIM.2012.6360147

Van Den Bergh, F. (2001). *An Analysis of Particle Swarm Optimizers*. PSO.

Van der Merwe, D. W., & Engelbrecht, A. P. (2003). Data clustering using particle swarm optimization. In Proceedings of the 2003 Congress on Evolutionary Computation CEC'03 (Vol. 1, pp. 215-220). IEEE. doi:10.1109/CEC.2003.1299577

Van Rijsbergen, C. J. (1979). Information retrieval. University Of Glasgow.

Wu Sen, C. H., Chen Hong, C. H., & Feng Xiaodong, F. X. (2013). Clustering algorithm for incomplete data sets with mixed numeric and categorical attributes. *International Journal of Database Theory and Application*, *6*(5), 95–104. doi:10.14257/ijdta.2013.6.5.09

Yang, X. S. (2010). A new metaheuristic bat-inspired algorithm. In Nature inspired cooperative strategies for optimization (NICSO 2010) (pp. 65-74). Springer Berlin Heidelberg. doi:10.1007/978-3-642-12538-6_6

Yang, X. S. (2010). Firefly algorithm, Levy flights and global optimization. In *Research and development in intelligent systems XXVI* (pp. 209–218). Springer London. doi:10.1007/978-1-84882-983-1_15

Yang, X. S., & Deb, S. (2009). Cuckoo search via Lévy flights. In *Proceedings of the World Congress on Nature and Biologically Inspired Computing NaBIC '09* (pp. 210-214). IEEE.

Yang, X. S., & Deb, S. (2010). Engineering optimisation by cuckoo search. *International Journal of Mathematical Modelling and Numerical Optimisation*, *1*(4), 330–343. doi:10.1504/IJMMNO.2010.035430

Yin, J., & Tan, Z. (2005). Clustering mixed type attributes in large dataset. In *Parallel and Distributed Processing and Applications* (pp. 655-661).

Zhang, C., Ouyang, D., & Ning, J. (2010). An artificial bee colony approach for clustering. *Expert Systems with Applications*, *37*(7), 4761–4767. doi:10.1016/j.eswa.2009.11.003

Zhao, X., & Lu, M. (2013). 3D Object Retrieval Based on PSO-K-Modes Method. *JSW*, *8*(4), 963–970. doi:10.4304/jsw.8.4.963-970

Zheng, Z., Gong, M., Ma, J., Jiao, L., & Wu, Q. (2010). Unsupervised evolutionary clustering algorithm for mixed type data. In *Proceedings of the 2010 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-8). IEEE. doi:10.1109/CEC.2010.5586136

KEY TERMS AND DEFINITIONS

Clustering: It is data mining technique to discover the hidden relationships between the data. It is the unsupervised learning technique and it groups the data objects without knowing class labels.

Data Mining: It is one of the steps in Knowledge Discovery in Databases (KDD). It discovers the interesting knowledge from large amount of data.

Optimization: These are the techniques to give the best possible solutions for the given objective problems. It minimizes the unfavorable solutions and maximizes the favorable solutions to the given problem.