Asian Research Consortium

Asian Journal of Research in Social Sciences and Humanities
Vol. 7, No. 2, February 2017, pp. 300-309.

ISSN 2249-7315
A Journal Indexed in Indian Citation Index
DOI NUMBER: 10.5958/2249-7315.2017.00092.2
Category:Science and Technology

Asian Journal of Research in Social Sciences and Humanities

www.aijsh.com

Cuckoo Search based K-Prototype Clustering Algorithm

K. Lakshmi*; Dr. N. Karthikeyani Visalakshi**; Dr. S. Shanthi***

*Research Scholar,
Anna University,
Chennai, Tamilnadu, India.
**Assistant Professor,
Department of Computer Science,
N. K. R. Government Arts College for Women,
Namakkal, Tamilnadu, India.
***Assistant Professor,
Department of Computer Applications,
Kongu Engineering College,
Perundurai, Tamilnadu, India.

Abstract

Data mining means extracting knowledge from big amount of data. Data mining techniques include clustering, Association, classification, Prediction, etc. Clustering is an unsupervised technique that is useful for finding groups in data. Most traditional clustering algorithms are limited to handling datasets that contain either numeric or categorical attributes. However, datasets with mixed numeric and categorical types of attributes are common in real life data mining applications. The K-Prototype clustering algorithm is one of the most important algorithms for clustering this type of data. This algorithm produces the locally optimal solution that is dependent on the initial prototype selection. This paper presents a new algorithm for data clustering based on K-Prototype and cuckoo search optimization to attain the global optimization.

Keywords: clustering, K-Means, K-Modes, K-Prototype Algorithm, Cuckoo Search Optimization Algorithm.

1. Introduction

Data clustering is the process of grouping the similar data into a number of clusters. Clustering algorithms plays important role in wide variety of data mining applications. Cluster is a collection of data objects that are similar to one another within the cluster and dissimilar to objects in other clusters (Han J. & Kamber M., 2006).

Clustering algorithms can be classified into partitional, Hierarchical, Density-based clustering and Grid-based. The simplest and most fundamental version of cluster analysis is partitioning, which arrange the data objects into several number of exclusive groups or clusters. A hierarchical clustering method works by grouping data objects into a hierarchy or “tree” of clusters.

K-means is one of the most widely used partitional clustering methods. This algorithm starts with randomly choose the K-objects from the dataset as a initial cluster centroids. Each data object is assigned to the nearest cluster centroid and each centroid is updated as the mean the all the data objects assigned to it. These two steps are repeated until no change in the clustership of the data objects. The disadvantage of K-Means is it handles only the numeric data.

K-Modes is the extension of K-Means clustering algorithm to cluster the categorical data objects (Huang, 1998). K-modes algorithm alter mean of cluster with mode and the mode is obtained by using simple matching dissimilarity obtained from the categorical data. The updated modes are obtained through the frequency based method.

(Huang, 1997) proposed the K-Prototypes algorithm which is based on the K-Means but removes the numeric data limitation and preserves the efficiency. This algorithm integrates the K-Means and K-Modes clustering algorithms to cluster the data objects with mixed numeric and categorical values. The initial centroids in these algorithms are selected as a random one. It may leads to different clustering results and falling into local optima. To avoid this inconvenience of K-Prototype clustering algorithm, this paper proposes a new metaheuristic approach and it is mainly based on the cuckoo search (CS) algorithm.

Cuckoo search (CS) is one of the latest nature-inspired metaheuristic optimization algorithm (Yang, X.-S., Deb, S, 2009). The advantages of CS include, its global search uses Lévy flights or process (Yang, X.-S., Deb, S, 2010). Cuckoo search is based on the interesting breeding behaviour such as brood parasitism of certain species of cuckoos and typical characteristics of Lévy flights. The CS is generic and robust for many optimization problems(Jothi, R. & Vigneshwaran, A, 2012) (Noghrehabadi, A. et al 2011). It is a population based and this algorithm overcomes the problem of local optimum to global one.

The remaining of this paper is organized as follows: In section 2, related works is presented. In section 3, introduction to cluster analysis is given. In section 4, we describe the basics of K-Prototype and cuckoo search algorithm. The proposed approach for data clustering is explained in section 5. Experimentation and comparisons are provided in Section 6. Finally, conclusions are drawn in Section 7.

2. Related Works

This section review the various algorithms proposed for clustering mixed numeric and categorical data, cuckoo search based K-Means clustering algorithm.

(Huang Z., 1998) proposed two algorithms, K-Means clustering algorithm to cluster the categorical data and combine the K-Means and K-Modes clustering algorithms to cluster the mixed numeric and categorical data. Also apply these algorithms to cluster the large data sets.

(Huang Z., 1997) proposed a way to dynamic updating of the K-Prototypes to maximize the intra cluster similarity of objects.

(Amir Ahmad & Lipika Dey, 2007) proposed the modified k-mean clustering algorithm for mixed datasets. Also introduce the new distance measure for categorical attributes.

A fuzzy K-Prototype clustering algorithm for mixed numeric and categorical data are proposed by (Jinchao Ji, et al, 2012). 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, et al, 2013). 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, 2014) 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.

(Tang, Rui, et al. 2012) proposed the way to integrating the nature inspired optimization algorithms to K-Means clustering algorithm.

(K. Arun Prabha, & N. Karthikeyani Visalakshi, 2015) proposed the Particle Swarm Optimization based k-prototype clustering algorithm to obtain the global optimum solution.

(Pham, Duc-Truong, et al, 2011) proposed the new algorithm called RANKPRO (Random Search with K-Prototypes algorithm), combines the advantages of a recently introduced population-based optimization algorithm called the bees algorithm (BA) and k-prototypes algorithm.

(Saida, et al, 2014) proposed the new algorithm for data clustering based on the cuckoo search optimization. The Cuckoo Search is to find K centroids of clusters which minimize the SSE.

(P. Manikandan & S. Selvarajan, 2014), proposes new approaches for using Cuckoo Search Algorithm (CSA) to cluster data. It is shown how Cuckoo Search Algorithm can be used to find the optimally clustering N object into K clusters. (Zhao, Jie, et al, 2014) proposed the Improved Cuckoo Search (ICS) algorithm for clustering

3. Cluster Analysis

Cluster analysis or simply clustering is the process of partitioning a set of data objects into cluster, such that objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. The set of clusters resulting from a cluster analysis can be referred to as a clustering. In this context, different clustering methods may generate different clustering on the same data set. The partitioning is not performed by humans, but by the clustering algorithm. Hence, clustering is useful in that it can lead to the discovery of previously unknown groups within the data.

Cluster analysis has been widely used in many applications such as business intelligence, image pattern recognition, Web search, biology, and security.

Clusters are formed by finding the distance between data objects and centroids. The quality of cluster Ci can be measured by the within cluster variation, which is the sum of squared error between all objects in Cluster and the centroid, defined as

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, c_i)^2,$$

where E is the sum of the squared error for all objects in the data set; **p** is the point in space representing a given object; and **ci** is the centroid of cluster Ci

4. Background

i. K-Prototype Clustering Algorithm

The K-Prototype clustering algorithm integrates the K-Means and K-Modes clustering algorithms. This algorithm is more useful for clustering mixed numeric and categorical data objects. According to the Huang, the cost function for mixed data objects is:

$$E = \sum_{l=1}^{k} \sum_{i=1}^{n} v_{il} d(X_i, Q_l)$$

Q_l is the prototype of the cluster l, v_{il} is an element of the partition matrix v_{nxk} and $d(X_i,Q_l)$ is the distance measure is calculated as follows:

$$dis(X_i, Q_l) = \sum_{j=1}^{p}(X_{ij}^r - Q_{lj}^r) + \gamma_l \sum_{j=p+1}^{m} \delta(x_{ij}^c, q_{lj}^c)$$

X^r_{ij}-Q^r_{lj} is the squared Euclidean distance measure for numeric attributes.

$\delta(X^c_{ij}$-$Q^c_{lj})$ is the simple matching dissimilarity measure on the categorical attributes.

γ_l is the weight for categorical attributes.

Algorithm: k-prototype Clustering Algorithm

Step 1. Select K initial prototypes from the data set X

Step 2. Allocate each data object in X to cluster whose prototype is nearest to it.

Step 3. Update the prototype of the cluster

Step 4. Retest the similarity of data objects against the updated prototype. If the data object is found that its nearest prototype belongs to another cluster rather than current one, reassign the data object to that cluster and update the prototype of both clusters.

Step 5. Repeat Step 3 until no data objects have change the clusters.

Step 6. Terminate the algorithm

ii. Cuckoo Search Algorithm

CS is based on the brood parasitism of some cuckoo species. In addition, this algorithm is enhanced by the Lévy flights. CS is potentially far more efficient than PSO, genetic algorithms, and other algorithms.

The standard CS, here we use the following three idealized rules:

- Each cuckoo lays one egg at a time and dumps it in a randomly chosen nest.
- The best nests with high-quality eggs will be carried over to the next generations.
- The number of available host nests is fixed, and the egg laid by a cuckoo is discovered by the host bird with a probability pa $\in (0, 1)$. In this case, the host bird can either get rid of the egg or simply abandon the nest and build a completely new nest.

Based on these three rules, the basic steps of the Cuckoo Search (CS) can be summarized by the pseudo code shown below:

```
begin
Objective function f(x), x = (x1, ..., xd)T
Generate initial population of n host nests xi (i = 1, 2, ..., n)
while (t <MaxGeneration) or (stop criterion)
Get a cuckoo randomly by Lévy flights
Evaluate its quality/fitness Fi
Choose a nest among n (say, j) randomly
if (Fi > Fj),
Replace j by the new solution;
end
A fraction (pa) of worse nests are abandoned and new ones are built;
Keep the best solutions (or nests with quality solutions);
Rank the solutions and find the current best
end while
Post process results and visualization
End
```

5. Proposed Algorithm

Cuckoo search based K-Prototype Clustering Algorithm

Step 1. Initialize the population of the host nests, N and related parameters

Step 2. Calculate the fitness of these population and find the best solution

Step 3. While t<MaxGen

i. Generate the new population with the cuckoo search

ii. Calculate the fitness of these new solutions

iii. Compare the new population with the old population, if new population is better than old one, replace the old with new population.

iv. Generate the fraction of new populations to replace the worst nests

v. Compare these solutions with old population. If the new population is better than old one, replace the old with new population.

vi. Find the best population.

Step 4. Output the best nest and fitness.

Step 5. Initialize the cluster centroids for K-Prototype with the best nest position particles of cuckoo search.

Step 6. Assign each particle of the population to closest centroid cluster of k-prototype.

Step 7. Recalculate the cluster centroid of k-prototype.

6. Results and Discussions

To test the validity and efficiency of the proposed algorithm, we have selected some of the datasets from UCI machine repository (Lichman, M. 2013). The dataset details are given in Table 1.

The performance of K-Prototype, PSO-based K-Prototype and Cuckoo search based K-Prototype algorithms are measured using external measures, Rand Index, F-Measure, Jaccard Index and Entropy. All these measures had the values between 0and 1.

Table 1: Dataset Description

S.No	Dataset	No.of Attributes	No. of Classes	No.of Instances
1.	Hepatitis	19	2	155
2.	Wine	13	3	178
3.	Bupa	6	2	345
4.	Satellite Image	36	7	6435
5.	Dermatology	34	6	366

The results of Cuckoo search based K-Prototype clustering algorithm compared with K-Prototype and PSO based K-Prototype are given in Table 2 to Table 6. It shows that Cuckoo search based K-Prototype gives the better results when compared with K-Prototype and PSO based K-Prototype clustering algorithms

Table 2: Comparative Analysis of Rand Index of Dataset

S.No	Dataset	k-prototype	PSO based k-prototype	Cuckoo search based k-prototype
1.	Hepatitis	0.6171	0.6723	0.7275
2.	Wine	0.4822	0.4998	0.5174
3.	Bupa	0.6853	0.6971	0.7089
4.	Satellite Image	0.5144	0.5312	0.548
5.	Dermatology	0.7029	0.7429	0.7829

Table 3: Comparative Analysis based on Jaccard Index

S.No	Dataset	K-Prototype	PSO based K-Prototype	Cuckoo search based K-Prototype
1.	Hepatitis	0.6099	0.6723	0.7347
2.	Wine	0.3870	0.3950	0.403
3.	Bupa	0.5240	0.5326	0.5412
4.	Satellite Image	0.3959	0.4141	0.4323
5.	Dermatology	0.5505	0.5705	0.5905

Table 4: Comparative Analysis based on F-measure

S.No	Dataset	K-Prototype	PSO based K-Prototype	Cuckoo search based K-Prototype
1.	Hepatitis	0.7217	0.7521	0.7825
2.	Wine	0.5615	0.5752	0.5889
3.	Bupa	0.8039	0.8229	0.8419
4.	Satellite Image	0.6041	0.6261	0.6481
5.	Dermatology	0.8197	0.8387	0.8577

Table 5: Comparative Analysis based on Entropy

S.No	Dataset	K-Prototype	PSO based K-Prototype	Cuckoo search based K-Prototype
1.	Hepatitis	0.3625	0.3421	0.3217
2.	Wine	0.6635	0.6231	0.5827
3.	Bupa	0.4896	0.5725	0.6554
4.	Satellite Image	0.6090	0.5961	0.5832
5.	Dermatology	0.4735	0.4635	0.4535

7. Conclusion

This paper proposed Cuckoo search based k-prototype clustering algorithm by combining cuckoo search with k-prototype to obtain the global optimum solution. The proposed algorithm has been tested with five different bench mark datasets which consists of both numeric and categorical attributes. It is proved that the performance of the proposed algorithm is better than the K-Prototype and PSO based K-Prototypes clustering algorithms (K. Arun Prabha, & N. Karthikeyani Visalakshi, 2015).

References

Amir Ahmad and Lipika Dey, A k-mean Clustering algorithm for Mixed Data with Numeric and Categorical Attributes, Data & Knowledge Engineering,63,503-527 (2007)

Arun Prabha K and Karthikeyani Visalakshi N., Particle Swarm Optimization based K-Prototype Clustering Algorithm, IOSR Journal of Computer Engineering 17 (2), 56-61, 2015

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

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)

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)

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

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)

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)

Jothi, R., Vigneshwaran, A.: An Optimal Job Scheduling in Grid Using Cuckoo Algorithm. International Journal of Computer Science and Telecommunications 3(2), 65–69 (2012)

Manikandan, P., and S. Selvarajan. "Data clustering using cuckoo search algorithm (CSA)." Proceedings of the Second International Conference on Soft Computing for Problem Solving (SocProS 2012), December 28-30, 2012. Springer India, 2014.

Noghrehabadi, A., Ghalambaz, M., Ghalambaz, M., Vosough, A.: A hybrid Power Series – Cuckoo Search Optimization Algorithm to Electrostatic Deflection of Micro Fixed-fixed Actuators. International Journal of Multidisciplinary Sciences and Engineering 2(4), 22–26 (2011)

Pham, Duc-Truong, Maria M. Suarez-Alvarez, and Yuriy I. Prostov. "Random search with k-prototypes algorithm for clustering mixed datasets." Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. Vol. 467. No. 2132. The Royal Society, 2011.

Saida, Ishak Boushaki, Kamel Nadjet, and Bendjeghaba Omar. "A new algorithm for data clustering based on cuckoo search optimization." Genetic and Evolutionary Computing. Springer International Publishing, 2014. 55-64.

Tang, Rui, et al. "Integrating nature-inspired optimization algorithms to K-means clustering." Digital Information Management (ICDIM), 2012 Seventh International Conference on. IEEE, 2012.

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Yang, X.-S., Deb, S.: Cuckoo Search via Levy Flights. In: Proc. of World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, pp. 210–214. IEEE Publications, USA (2009)

Yang, X.-S., Deb, S.: Engineering Optimisation by Cuckoo Search. International Journal of Mathematical Modelling and Numerical Optimisation 1(4-30), 330–343 (2010)

Zhao, Jie, et al. "Clustering Using Improved Cuckoo Search Algorithm." International Conference in Swarm Intelligence. Springer International Publishing, 2014.