INTERNATIONAL RESEARCH JOURNAL OF ENGINEERING AND TECHNOLOGY (IRJET) E-ISSN: 2395-0056
VOLUME: 06 ISSUE: 02 | FEB 2019 WWW.IRJET.NET P-ISSN: 2395-0072

# A LION OPTIMIZATION BASED K-PROTOTYPE CLUSTERING ALGORITHM FOR MIXED DATA

**G.S. Nithya[1], K. Arun Prabha[2]**

*[1]Research Scholar, Department of Computer Science, Vellalar College for Women, Erode, Tamil Nadu, India*
*[2]Head and Assistant Professor, Department of Computer Technology (IT & CT), Vellalar College for Women, Erode, Tamil Nadu, India*

---

**Abstract--** *Data Mining is used to extract information from huge set of data. Clustering is the task of grouping a set of objects. The K-Means clustering is only used for numeric data which has local optima. The K-Modes extends to the K-Means when the domain is categorical. The K-Prototype algorithm is one of the most important algorithms for clustering mixed type of data. This algorithm is very beneficial for clustering large data sets. Lion Optimization Algorithm is one of the simple optimization techniques, which can be effectively implemented to enhance the clustering results. It is useful for handling mixed data set. This leads a better optimization for calculating the centroid with the K-Prototype clustering algorithm. To overcome the issues in K-Prototype clustering algorithm the Lion optimization Algorithm is used. The proposed algorithm is implemented on standard benchmark dataset taken from UCI Machine Learning Repository. Through optimizing the K-Prototype clustering with Lion Optimization Algorithm have a better performance than K-Prototype clustering algorithm.*

**Keywords—** Data Mining, Clustering Model, K-Means cluster, K-Prototype cluster, Lion Optimization Algorithm

## 1. INTRODUCTION

Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD. It is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Clustering is an un-supervised learning. A cluster is a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering is also used to reduce the dimensionality of the data when you are dealing with a copious number of variables. The goal of clustering is to discover both the dense and the sparse regions in a dataset.

There are two types of clustering

- Hierarchical Clustering
- Partitional Clustering

### Hierarchical Clustering

The hierarchical clustering is an algorithm that groups similar objects into groups. This hierarchy of clusters is represented as a structure of a tree.

There are two types of hierarchical clustering, Divisive and Agglomerative. In top-down or divisive clustering method we assign all of the observations to a single cluster and then partition the cluster to two least similar cluster. In bottom-up or agglomerative clustering method we assign each observation to its own cluster.

### Partitional Clustering

The partitional clustering are clustering methods used to classify observations, within a data set, into multiple groups based on their similarity. The commonly used partitional clustering are K-Means clustering, K-Modes Clustering or PAM and CLARA algorithm. The K-Means clustering and K-Modes clustering is mainly used in the partitioning algorithm. The K-Means clustering represent each cluster by the center of gravity. The K-Modes clustering represents each cluster by the cluster located near center. The partitional clustering has three types of clustering

- K-Means Clustering
- K-Modes Clustering
- K-Prototype Clustering

The system is defined with the following specific objectives.

- The main objective in this work is to optimize the K-Prototype clustering algorithm using Lion Optimization Algorithm
- To overcome the K-Prototype algorithm a Lion Optimization Algorithm is introduced.

Five various datasets are taken from the UCI repository; the datasets are applied to Lion Optimization Algorithm and K-Prototype algorithm. The results are evaluated and the validity measure like F-Measure, Rand index, Jaccard index and Entropy are used.

## 2. RELATED WORK

**Zhexue Huang [1997],** analyzed the K-Prototype clustering algorithms for mixed data such as numeric and categorical data. The K-Means based methods have the efficiency of the large datasets and it has limited numeric value to be evaluated. In the research they have introduced a K-Prototypes algorithm based on the K-Means partitions to removes the numeric data limitation. A method was developed to dynamically update the K-Prototype in order to maximize the intra cluster similarity of objects. The decision tree induction algorithm is used for creating rules for clusters and to understand and identify interesting clusters.

**Ming-Yi Shih et al [2010]**, has proposed a two step method for clustering mixed categorical and numeric data. Clustering algorithm work effectively with pure numeric data or pure categorical data. But it has been work poorly with the mixed data such as numeric and categorical data. The two step clustering was used for the dissimilarity measures to deal with both categorical and numerical data. A two step method has been introduced for integrating hierarchal and partitioning clustering algorithm for the weakness of K-Means algorithm. They have proposed a new approach as a single clustering algorithm to explore the relationship among the categorical values and numeric values. The categorical values have converted into numerical values and the numeric values were applied for the data sets.

**Wu Sen. et al [2013]**, has proposed a K-Prototype clustering algorithm for incomplete datasets with mixed numeric and categorical attributes. The traditional K-Prototype algorithm is well versed in clustering data with mixed numeric and categorical attributes, while the completed data are limited. To handle incomplete dataset with missing values, an improved K-Prototype algorithm were proposed, which employs a new dissimilarity measure for incomplete dataset with mixed numeric and categorical attributes. A new approach was used to select K objects as the initial prototypes based on the nearest neighbors. To illustrate the accuracy of the established algorithm, traditional K-Prototype algorithm and K-Prototype employing the new dissimilarity measure were compared to the improved K-Prototype algorithm. The new dissimilarity measure computation takes into account missing data, with no need to impute missing data with means or modes before clustering, which decreases an estimation that might cause some error.

**Izhar Ahmad et al [2014]**, analyzed as K-Means and K-Prototype as performance analysis. The system design approach has been presented for the K-Means and K-Prototype performance analysis. The system architecture in the research have presented an detail discussion of the K-Means and K-Prototype algorithm to recommend efficient algorithm for outlier detection and other issues which are related to the database clustering. The original algorithm of the K-Means and K-Prototype algorithm does always the guarantee the accuracy of final clusters which are based on the selection of initial centroids. The proposed system architecture have utilize the complete unified solution for the K-Means and K-Prototype algorithms of performance analysis. The analysis have shown that the proposed system architecture procedures better clusters in less computation time as compared to the standard K-Means and K-Prototype algorithm.

**K.Arun Prabha et al [2015]**, proposed a new variant of binary Particle Swarm Optimization and K-Prototype algorithms to reach global optimal solution for clustering optimization problem. The comparative analysis of K-Prototype and PSO proved that Particle Swarm based on K-Prototype algorithm provides better performance than the traditional K-Modes and K-Prototype algorithms. 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. It is proved that the performance of the proposed algorithm is superior to the performance of conventional K-Modes and K-Prototype algorithms.

## 3. METHODOLOGY

### Lion Optimization based K-Prototype clustering algorithm

K-Prototype Clustering is an effective algorithm for clustering mixed type of 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, Lion Optimization Algorithm and K-Prototype algorithms are combined. In this method, the process is initialized with a group of random population (Lion). The Lion Optimization based K-Prototype algorithms consists of the following steps:

Algorithm: Lion Optimization based K-Prototype Clustering Algorithm

Input: Mixed data, Parameters $N_{pop}$ as population, %N as nomad lion, %S as female lion, iteration K, Number of clusters K, assign threshold σ.

Output: K cluster

Procedure

Step 1: Separate the categorical data and numeric data.

Step 2: Change the categorical into numeric data using K-Mode by

$$P(W,Q) = \sum_{l=1}^{K}\sum_{i=1}^{n} w_i d_i d_{sim}(x_i, q_l) \quad (1)$$

Step 3: Preprocess the data.

Step 4: Generate initial population of lion $N_{pop}$.

Step 5: Initiate the nomad and pride

i) Randomly select %N as nomad lion and partition the remaining lions into pride.

ii) In each pride %S of entire population is female lion.

Step 6: Each gender of the nomad lion are sorted based on the fitness value by

$$f_i = \sum_{i=1}^{c}\sum_{d_m \epsilon c_{cen}} ||d_m - c_{cen}||^2 \quad (2)$$

Step 7: Arrange the fitness values from minimum to the maximum values.

Step 8: Choose the minimum value from the fitness value.

Step 9: Repeat Step 4 to Step 7 until the cluster of centroid is selected.

Step 10: Apply the fitness value in K-Prototype clustering algorithm.

Step 11: Calculate the Euclidean distance by using

$$d(x,y)=\sqrt{\sum_{i-1}^{n}(x_i - y_i)^2} \quad (3)$$

Step 12: Recalculate the centroid.

Step 13: Repeat Step 10 to Step 12 until the clustering are not changed.

## 4. RESULTS AND DISCUSSION

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. The details of the data sets are given in the following Table-1.

The performance of K-Means, K-Prototype and Lion Optimized Algorithm based K-Prototype clustering algorithm is measured in terms of four external validity measures namely F-Measure, Rand Index, Jaccard Index and Entropy. The external validity measures test the quality of clusters by comparing the results of clustering. 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 Instance | No.of Attributes | No. of Classes |
|---|---|---|---|---|
| 1 | Australian Credit Approval | 690 | 14 | 2 |
| 2 | German Credit Approval | 1000 | 20 | 2 |
| 3 | Hepatitis | 155 | 19 | 2 |
| 4 | Post Operative Patient | 90 | 8 | 3 |
| 5 | Stat Log Heart | 270 | 13 | 2 |

According to Rand Index, the performance of Lion Optimization Algorithm based K-Prototype clustering algorithm yields consistent and improved results than K-Means and K-Prototype Algorithm in almost all datasets. From the Table-2 and Figure-1 it is observed that Lion Optimization Algorithm based K-Prototype clustering 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: Comparitive analysis based on Rand Index**

| S. No | Dataset | K-Means | K-Prototype | Lion Optimization based K-Prototype Clustering Algorithm |
|---|---|---|---|---|
| 1 | Australian Credit Approval | 0.56 | 0.58 | 0.62 |
| 2 | German Credit Data | 0.53 | 0.55 | 0.57 |
| 3 | Hepatitis | 0.72 | 0.73 | 0.74 |
| 4 | Post Operative Patient | 0.41 | 0.45 | 0.47 |
| 5 | Stat log Heart | 0.66 | 0.68 | 0.72 |

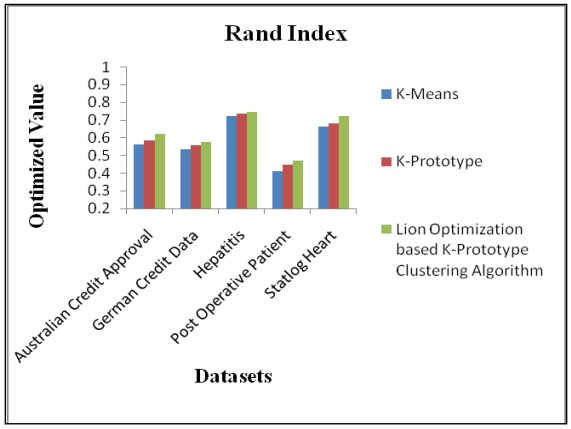


**Figure-1 Analysis Based on Rand Index**

From the Table-3, based on Jaccard Index, the performance of Lion Optimized Algorithm based K-Prototype clustering algorithm yields consistent and better results for data sets. K-Means and K-Prototype algorithm in almost all datasets.

**Table-3:Comparitive analysis based on Jaccard Index**

| S. No | Dataset | K-Means | K-Prototype | Lion Optimization based K-Prototype Clustering Algorithm |
|---|---|---|---|---|
| 1 | Australian Credit Approval | 0.43 | 0.45 | 0.52 |
| 2 | German Credit Data | 0.45 | 0.46 | 0.48 |
| 3 | Hepatitis | 0.62 | 0.63 | 0.65 |
| 4 | Post Operative Patient | 0.33 | 0.35 | 0.38 |
| 5 | Stat log Heart | 0.54 | 0.56 | 0.70 |

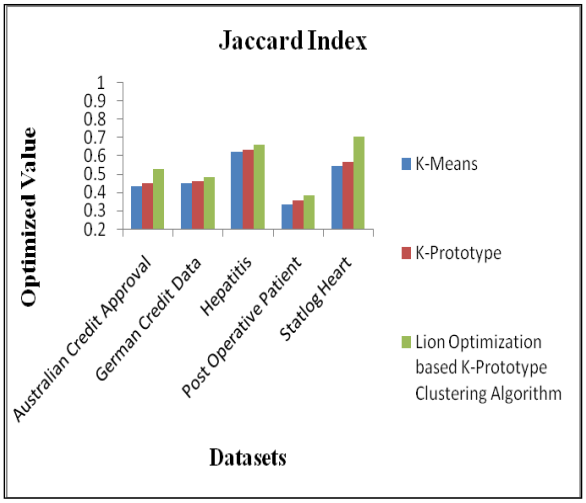


**Figure 2 Analysis Based on Jaccard Index**

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 Lion Optimization Algorithm based K-Prototype Clustering Algorithm is appreciable

than K-Mean and K-Prototype algorithm for all datasets represented in Table 4.

**Table-4: Comparitive analysis based on F-Measure**

| S. No | Dataset | K-Means | K-Prototype | Lion Optimization based K-Prototype Clustering Algorithm |
|---|---|---|---|---|
| 1 | Australian Credit Approval | 0.76 | 0.83 | 0.86 |
| 2 | German Credit Data | 0.76 | 0.82 | 0.84 |
| 3 | Hepatitis | 0.78 | 0.84 | 0.87 |
| 4 | Post Operative Patient | 0.79 | 0.85 | 0.86 |
| 5 | Stat log Heart | 0.85 | 0.87 | 0.88 |

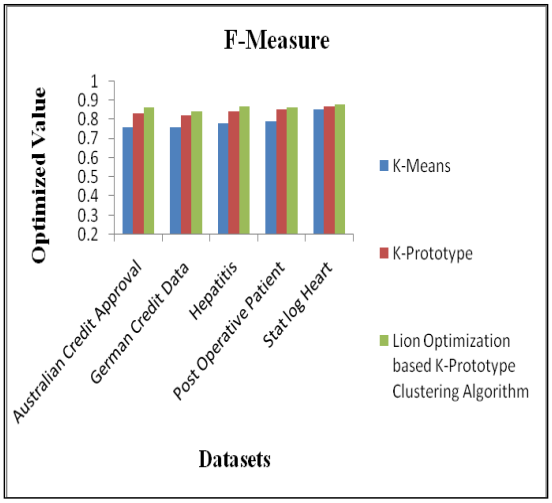


**Figure 3 Analysis Based on F-Measure**

The decrease in the values of Entropy measure proves the better performance. Based on that the performance of Lion Optimized Algorithm based K-Prototype clustering algorithm based on Entropy is highly significant than K-Mean and K-Prototype for all dataset represented in Table-5.

**Table-5: Comparitive analysis based on Entropy**

| S. No | Dataset | K-Means | K-Prototype | Lion Optimization based K-Prototype Clustering Algorithm |
|---|---|---|---|---|
| 1 | Australian Credit Approval | 0.51 | 0.51 | 0.50 |
| 2 | German Credit Data | 0.45 | 0.43 | 0.45 |
| 3 | Hepatitis | 0.52 | 0.52 | 0.52 |
| 4 | Post Operative Patient | 0.42 | 0.41 | 0.40 |
| 5 | Stat log Heart | 0.45 | 0.44 | 0.43 |

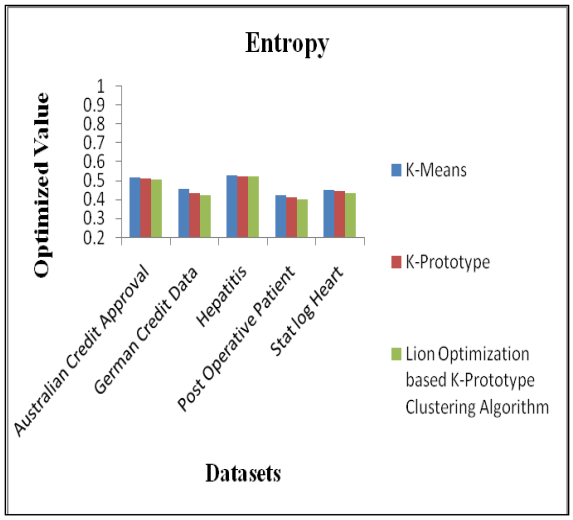


**Figure.4 Analysis Based on Entropy**

## 5. CONCLUSION

This paper proposed Lion Optimization based K-Prototype Clustering algorithm by incorporating the benefit of Lion Optimization algorithm with the existing K-Prototype algorithm, to reach the global optimum cluster solution. The proposed algorithm has been tested on the five benchmark datasets 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-Means and K-Prototype clustering algorithms.

**REFERENCES**

[1] Arun Prabha .K, N. Karthi Keyani Visalakshi,"Particle Swarm Optimization based K-Prototype Clustering Algorithm" IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, ISSN: 2278-8727, Vol 17, pp. 56-62, [2015].

[2] Fengmei W and H. Lixia, "A Missing Data Imputation Method Based on Neighbor Rules", Computer Engineering, vol. 38, no. 21, [2012].

[3] Izhar Ahmad,"K-Mean and K-Prototype Algorithms Performance Analysis", American Review of Mathematics and Statistics, ISSN 2374-2348, Vol. 2, pp. 95-109, [2014].

[4] Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han,Zhe Wang,"A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", Elsevier , ISSN 0950-7051,Vol. 30,[2012].

[5] Li Xinwu,"A New Text Clustering Algorithm Based on Improved K-Means", Journal of Software, Vol. 7, doi:10.4304/jsw.7.1.95-101, [2012].

[6] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai," A Two-Step Method for Clustering Mixed Categroical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, pp. 11-19, [2010].

[7] Maziar Yazdani, Fariborz Jolai,"Lion Optimization Algorithm (LOA): A Nature-Inspired Metaheuristic Algorithm", Elsevier, Journal of Computational Engineering, [2015].

[8] Wu Sen, Chen Hong and Feng Xiaodong. "Clustering Algorithm for Incomplete Data Sets with Mixed Numeric and Categorical Attributes", International Journal of Database Theory and Application, ISSN:2005-4270 IJDTA, Vol.6, pp.95-104, [2013].

[9] Zhexue Huang,"Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values", Data Mining and Knowledge Discovery 2,pp. 283-304, [1998]

[10] Zhexue Huang, "Clustering Large Data Sets with Mixed Numeric and Categorical Values", CSIRO Mathematical and Information Sciences, Conference, pp.21-34, [1997]