IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-5, Issue-6 E-ISSN: 2347-2693

# Optimized K-Mode Algorithm Using Harmonic Technique

**Manisha Goyal[1*], Shruti Aggarwal[2]**

[1*]Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India
[2]Department of Computer Science and Engineering, Sri Guru Granth Sahib World University, Fatehgarh Sahib, India

**Corresponding Author: manisha93goyal@gmail.com*

**Available online at: www.ijcseonline.org**

Received: 25/May/2017, Revised: 02/Jun/2017, Accepted: 20/Jun/2017, Published: 30/Jun/2017

***Abstract***— Data Mining is the extraction of useful information from a huge amount of datasets. As one of the most important tasks in data mining, clustering aims to group a set of objects such that the objects within the same cluster are more similar to each other than to the objects in another cluster. An extension of the K-Means Algorithm, K-Mode Algorithm, is partitioning based clustering algorithm does not guarantee for the optimal solution. To overcome this problem, entropy based similarity coefficient was introduced in order to find good initial center points and the accurate result of the clusters were obtained. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. In this paper, Harmonic K-Mode Algorithm is proposed that reduces the computation time and improves the accuracy for cluster generation. The experimental result shows that the proposed algorithm gives better results than the existing algorithms.

***Keywords***— Data Mining, Clustering, K-Means Algorithm, K-Mode Algorithm

## I. Introduction

Data Mining is the process to extract previously unknown, valid patterns and relationships that provide useful information and can be used to make certain decisions [2]. The most diverse characteristic of data mining is that it deals with very large and complex data sets. The datasets to be mined contain millions of objects described by tens, hundreds or even thousands of various types of attributes or variables. Major components of data mining technology have been under development, such as statistics, artificial intelligence and machine learning in research areas [1]. In data mining, one needs to be concentrating on the cleansing of the data for the further processing to make it feasible. This process is known as the noise elimination or noise reduction [3]. The data mining operations and algorithms are required to deal with different types of attributes. In this sophisticated data analysis tools are used along with visualization techniques to segment the data [2]. The various application areas of data mining are marketing, education, banking, insurance, transportation, healthcare, finance etc.

Rest of the paper is organized as follows, section I contains the introduction of the data mining, section II contains the clustering using k-mode along with its algorithm, section III contains literature survey, section IV contains the proposed work, section V contains the experimental results and section VI contains the concludes the research work with future scope.

## II. Clustering using K-Mode

Clustering is a form of unsupervised learning that means how data should be grouped with the data objects (similar types) together will be not known in advance. It is the technique which defines classes and put objects in one class which are related to them, which means to put the objects into one group having similar properties and objects having dissimilar properties into another group [3].

The clustering algorithms can be generally classified into five categories: Hierarchical Based Clustering, Partitioning Based Clustering, Density Based Clustering, Grid Based Clustering and Model Based Clustering. Partitioning Based Clustering is the popular approach of clustering, which transfer objects by moving them from one cluster to another cluster starting from a certain point. The amount of clusters for this technique should be predefined and the algorithms used in this approach are K-Means Algorithm, K-Medoid Algorithm, K-Nearest Neighbour Algorithm etc.

K-Means Algorithm is popular unsupervised learning algorithm and is a partitioning based algorithm for clustering. K-Means clustering approach creates clusters of the same type of data according to their closeness to each other based on the Euclidean distance [4]. It intends to partition the objects into a number of clusters in which each object belongs to the cluster with the nearest mean [5].

K-Mode Algorithm is an extension of K-Means Algorithm and is the partitioning based clustering algorithm. It uses simple matching dissimilarity function instead of using Euclidean distance. Modes are used to represent centroids and a frequency based method is used to find the centroids in each iteration of the algorithm [13].

**Algorithm for K-Mode Clustering:**

The steps of the K-Mode Algorithm are as follow:

1. *Select k unique objects as the initial cluster centers randomly.*
2. *Calculate the distances between each object and the cluster mode.*
3. *Assign the object to the cluster whose center has the shortest distance to the object.*
4. *Repeat this step until all objects are assigned to clusters.*
5. *Select a new mode for each cluster and compare it with the previous mode. If different, repeat steps 2 to 4 until the criteria are fulfilled.*

The numbers of initial cluster centers are chosen randomly. This algorithm creates clusters of the same type of data according to their closeness to each other based on the simple matching dissimilarity function. It intends to partition the objects into a number of clusters in which each object belongs to the cluster with the nearest mode. It is famous for its simplicity and speed.

## III. Literature Survey

K-Mode algorithm is an extension of K-Means Algorithm, as the mode is calculated instead of calculating the mean value in order to find the accurate clusters to avoid overlapping.

One of the K-Mode Algorithm is a K-Prototype Algorithm [6] integrates the dissimilarity measure in the K-Means and K-Mode algorithms for clustering objects having mixed numeric and categorical and results in allocation of less similar objects in a cluster.

Another K-Mode Algorithm is Iterative K-Mode [7] introduced an initialization method based on Bradley's iterative initial-point refinement algorithm to the K-Modes clustering and results in accurate number of clusters.

Another one of the K-Mode Algorithm is COOLCAT Algorithm [8] introduced an algorithm which is able to deal with clustering of data streams and is based on the notion of entropy. It depends on an input parameter that represents the size of the small cluster.

One more algorithm is Modified K-Mode [9] proposed a new notion of cluster centres and dissimilarity measure is introduced and showed better result over the K-Mode Algorithm.

The other k-mode algorithm is Distance based k-mode [10] proposed an initialization method for categorical data and the distance between objects was calculated based on the frequency of attribute values, calculates the densities of all the objects of categorical data and the process was limited to sub-sample dataset.

One more K-Mode Algorithm is a DVD based K-Mode Algorithm [11] suggested a new measure called Domain Value Dissimilarity. The information about the distribution of data correlated to each categorical value was used to define the dissimilarity measure.

Another K-Mode Algorithm is DILCA Algorithm [12] proposed a method called Distance learning for Categorical Attributes. The distance between two values of a categorical attribute was determined by the way in which the value of the other attributes was distributed in the dataset.

DISC algorithm [13] is another one K-Mode Algorithm that suggested the method Data-Intensive Similarity Measure for Categorical Data. This measure didn't require any domain knowledge to understand the dataset.

Biological and Genetic taxonomy information based k-mode [14] is the other K-Mode Algorithm that proposed a new dissimilarity measure based on the idea of biological and genetic taxonomy and rough membership function and improved the accuracy of the clusters.

K-Mode based upon the unified similarity metrics algorithm [15] proposed a penalized competitive learning algorithm and this algorithm required some initial value of the number of clusters which should be greater than the original value of the number of clusters. The resulting clusters are more accurate than the original K-Mode Algorithm.

Cluster centre initialization based k-mode algorithm [16] is one of the K-Mode Algorithm that introduced some objects which are very similar to each other and have same cluster membership irrespective of the choice of initial cluster centres and generate accurate clusters using prominent attributes.

Entropy based similarity coefficient k-mode algorithm (EC K-Mode) [17] is another K-Mode Algorithm that improved the cluster accuracy and analysed the time complexity while retaining the scalability of the K-Mode Algorithm.

In K-mode with Maritime clustering [18] systematized the features of preconditions specific to Maritime clustering, which the increase of Productivity, Innovations and Competitiveness which is also the other K-Mode Algorithm.

**Entropy based Similarity Coefficient K-Mode Algorithm**

The steps of Entropy based Similarity Coefficient K-Mode Algorithm are as follow:

1. Initialize the variable old modes as a k × m empty array
2. Randomly k different data points from dataset as initial modes will be chosen and assign [$m_1$, $m_2$, ……., $m_k$] to k × m array variable new modes.

```
3.  for i=1 to N do
4.  for l=1 to k do
5.  Calculate the similarity between ith data point and lth mode vector using similarity coefficient.
6.  Assign that data point to appropriate cluster whose cluster mode vector is closer to it.
7.  Update mode vector of corresponding cluster and find the distribution of mode categories between clusters.
8.  end
9.  end
10. while old modes!= new modes
11. do old modes = new modes
12. for i=1 to N do
13. for l=1 to k do
14. Calculate the similarity between ith data point and lth mode vector using similarity coefficient.
15. Assign that data point to appropriate cluster whose cluster mode vector is closer to it.
16. Update mode vectors of corresponding two clusters and find the distribution of mode categories between clusters.
17. end
18. end
19. if old modes = new modes then
20. break
21. endif
22. end
```

The algorithm for the Entropy based similarity coefficient k-mode algorithm (EC K-Mode) was used to find the optimal centroids and used a similarity coefficient based on information entropy. This algorithm improves the accuracy of the clusters and the time complexity of this algorithm is O(n) which is linearly scalable to the k-mode algorithm. The execution time and accuracy of this algorithm is improved by our proposed algorithm.

## IV. Proposed Work

The k-mode algorithm is a partitioning based clustering algorithm and is an extension of K-means algorithm. It is linearly scalable with respect to the dataset size. It is especially sensitive to the selection of initial cluster centers and choosing the proper initial cluster centers is a key step for k-mode clustering and does not guarantee for the optimal solution. To overcome this problem, entropy based similarity coefficient was introduced in order to find good initial center points to obtain the accurate result of the clustering. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. Harmonic K-Mode Algorithm is proposed in this work that may reduce the computation time and improves the accuracy for cluster generation.

**Algorithm for Harmonic K-Mode Clustering**

The steps of the Harmonic K-Mode Algorithm are as follow:

```
1.  Initialize the variable old modes as an k × m empty array
2.  Randomly k different data points from dataset as initial modes will be chosen and assign [$m_1$, $m_2$,……., $m_k$] to k × m array variable new modes.
3.  for i=1 to N do
4.  for l=1 to k do
5.  Calculate the similarity between ith data point and lth mode vector using similarity coefficient.
6.  Assign that data point to appropriate cluster whose cluster mode vector is closer to it.
7.  Update mode vector of corresponding cluster and find the distribution of mode categories between clusters.
8.  end
9.  end
10. while old modes != new modes
11. do old modes = new modes
12. for i=1 to N do
13. for l=1 to k do
14. Initialization of harmony memory that consists of randomly generated solution to the optimization problem.
15. Improvise a new solution from the harmony memory.
16. Update the harmony memory.
17. Repetition of Steps 15 and 16 until the termination criterion is satisfied.
18. end
19. end
20. if old modes = new modes then
21. break
22. endif
23. end
```

The harmonic k-mode algorithm is hybrized to find the optimal value of the centroids so that strong clusters may be formed. After the initialization of the data points, it calculates the similarity of the data point and the mode using similarity coefficient. The data points will be assigned to the appropriate cluster. The solutions are stored in the harmony memory. It helps to find the optimal values of the centroids that gives better accuracy and reduces the computation time.

The basic design of the proposed algorithm shows that the dataset is selected to perform clustering. After selecting the datasets, three algorithms are used to perform clustering i.e. k-mode algorithm, k-mode with entropy based similarity coefficient for clustering and harmonic k-mode algorithm. The basic design of the proposed algorithm is shown in figure 1.

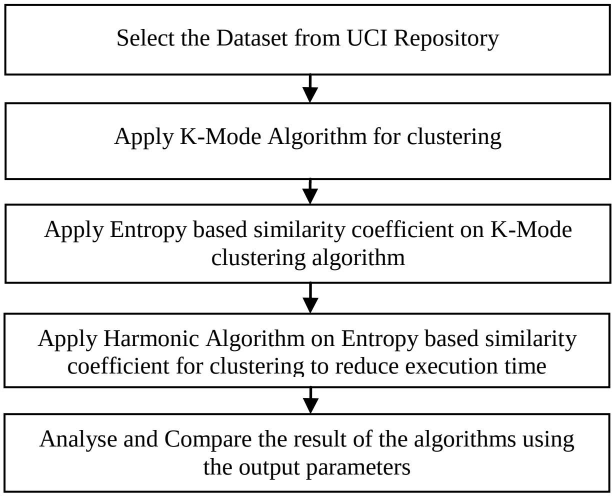

Select the Dataset from UCI Repository
Apply K-Mode Algorithm for clustering
Apply Entropy based similarity coefficient on K-Mode clustering algorithm
Apply Harmonic Algorithm on Entropy based similarity coefficient for clustering to reduce execution time
Analyse and Compare the result of the algorithms using the output parameters

Fig 1: Basic Design of the Proposed Algorithm

The datasets are selected from the UCI Repository. K-Mode and Entropy based similarity coefficient k-mode are the existing algorithms and Harmonic k-mode is the proposed algorithm. The results of the algorithms are compared using five output parameters. The datasets, output parameters and results of these algorithms are discussed in the next section.

## V. EXPERIMENTAL RESULTS

### A. Datasets

The datasets used in this work is taken from UCI Repository dataset, i.e. amazon book review, news aggregator, online retail, seed and wholesale consumer. The total number of instances is 1500 and attributes is 10000 in the amazon book review dataset. The total number of instances is 42293 and attributes is 5 in the news aggregator dataset. The total number of instances is 54190 and attributes is 8 in the online retail dataset. The total number of instances is 210 and attributes is 7 in the seed dataset. The total number of instances is 440 and attributes is 8 in the wholesale consumer dataset.

### B. Output Parameters

The result is compared using five parameters execution time, space complexity, accuracy, precision and recall.

- **True Positive Rate (TP):** A true positive test result is one that detects the condition when the condition is present.
- **True Negative Rate (TN):** A true negative test result is one that does not detect the condition when the condition is absent.
- **False Positive Rate (FP):** A false positive test result is one that detects the condition when the condition is absent.
- **False Negative Rate (FN):** A false negative test result is one that does not detect the condition when the condition is present.

Following output parameters are used to evaluate the performance of the algorithms:

i. **Execution Time:** The execution time is defined as the time spent by the system executing the task.

ii. **Space Complexity:** Space complexity is a measure of the amount of working storage an algorithm needs means how much memory is needed at any point in the algorithm.

iii. **Accuracy:** The Accuracy is the total number of module that is predicted correctly.

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \quad (1)$$

iv. **Precision:** Precision is the measure of exactness i.e. what percentage of tuples labeled as positive that are actually such.

$$\text{Precision} = \frac{(TP)}{(TP+FP)} \quad (2)$$

v. **Recall:** Recall is the measure of completeness i.e. what percentage of positive tuples did the classifier labelled as positive.

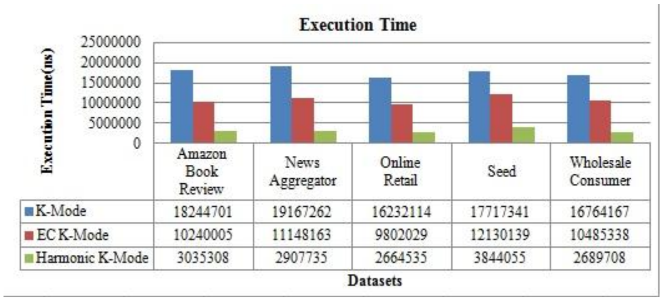
$$\text{Recall} = \frac{(TP)}{(TP+FN)} \quad (3)$$

### C. Results

The performance of the algorithms can be evaluated by using output parameters for the number of clusters 10. The comparisons of the results among the algorithms on different datasets are shown as follows:

#### i. Execution Time

The execution time is computed in nanoseconds for the algorithms applied to different datasets are shown in the Figure 2.



| | Amazon Book Review | News Aggregator | Online Retail | Seed | Wholesale Consumer |
|---|---|---|---|---|---|
| 🟦 K-Mode | 18244701 | 19167262 | 16232114 | 17717341 | 16764167 |
| 🟥 EC K-Mode | 10240005 | 11148163 | 9802029 | 12130139 | 10485338 |
| 🟩 Harmonic K-Mode | 3035308 | 2907735 | 2664535 | 3844055 | 2689708 |

Fig 2: Execution Time

The Figure 2 shows that the Harmonic K-Mode Algorithm has better execution time than K-Mode and EC K-Mode Algorithm.

#### ii. Space Complexity

The space complexity is computed in bytes for the algorithms applied to different datasets are shown in the Figure 3.

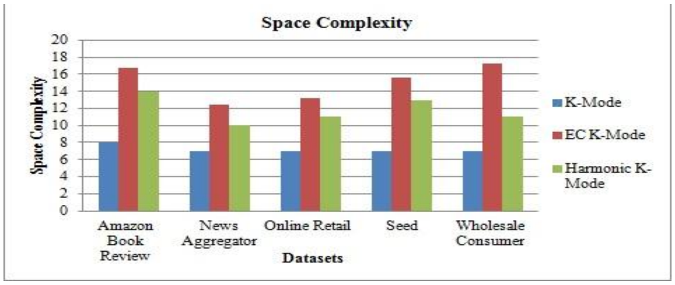


Fig 3: Space Complexity

The Figure 3 shows that the space complexity for Harmonic K-Mode algorithm is less than the EC K-Mode algorithm.

### iii. Accuracy

The graph of the accuracy of the algorithms for different datasets is shown in the Figure 4.

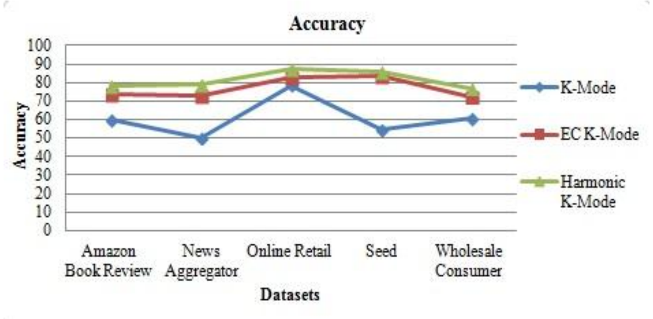


Fig 4: Accuracy

The Figure 4 shows that the Harmonic K-Mode Algorithm has the better accuracy than K-Mode and EC K-Mode Algorithms.

### iv. Precision

The comparison of the precision of the algorithms for different datasets is shown in the Table 1.

Table 1: Precision

| Dataset | K-Mode | EC K-Mode | Harmonic K-Mode |
|---|---|---|---|
| Amazon Book Review | 51 | 56 | 57.6 |
| News Aggregator | 51.75 | 67.8 | 68.6 |
| Online Retail | 54.3 | 59 | 64 |
| Seed | 46.5 | 56.2 | 73.52 |
| Wholesale Consumer | 57.7 | 65.4 | 60.87 |

The Table 1 shows that the precision for Harmonic K-Mode Algorithm is better than the existing algorithm, but EC K-Mode has more precision than K-Mode and Harmonic K-Mode due to the random selection of initial centroids in the wholesale consumer dataset.

### v. Recall

The comparison of the recall of the algorithms for different datasets is shown in the Table 2.

Table 2: Recall

| Dataset | K-Mode | EC K-Mode | Harmonic K-Mode |
|---|---|---|---|
| Amazon Book Review | 66.3 | 70.2 | 77 |
| News Aggregator | 54.3 | 59.4 | 62 |
| Online Retail | 66.75 | 75.64 | 80.64 |
| Seed | 64.56 | 69.7 | 74 |
| Wholesale Consumer | 55.2 | 62.1 | 53 |

The Table 2 shows that the recall for Harmonic K-Mode Algorithm is better than the existing algorithm, but EC K-Mode has more recall than K-Mode and Harmonic K-Mode due to the random selection of initial centroids in the wholesale consumer dataset.

## VI. Conclusion and Future Scope

The K-Mode Algorithm, an extension of K-Means Algorithm, is linearly scalable with respect to the dataset size. It is especially sensitive to the selection of initial cluster centers and choosing the proper initial cluster centers is a key step for k-mode clustering. It does not guarantee for the optimal solution. To overcome this problem, k-mode with entropy based similarity coefficient was introduced in order to find good initial center points and the accurate result of the clustering was to be obtained. The nature-inspired harmonic algorithm is hybridized to optimize the k-mode algorithm. Harmonic K-Mode Algorithm is the proposed algorithm that may reduce the computation time and improves the accuracy for cluster generation. Different parameters are used to analyse the result of optimized k-mode algorithm. The performance of the harmonic k-mode is evaluated in terms of execution time, space complexity, accuracy, precision and recall on the five datasets. The experimental results show that the proposed algorithm has better results than the existing algorithms.

For the future, the performance of the algorithms can be enhanced with no dependency on number of clusters that reduce the computation time for clusters generation by using some better optimization algorithms.

## References

[1]. V. Sawant, K. Shah, “*Performance Evaluation of Distributed Association Rule Mining Algorithms*”, 7th International Conference on Communication, Computing and Virtualization, Elsevier, Vol. 79, pp. 127-134, 2016.

[2]. J. Karimov, M. Ozbayoglu, “*Clustering Quality Improvement of k-means using a Hybrid Evolutionary Model*”, Conference Organized by Missouri University of Science and Technology, San Jose, California, Elsevier, Vol. 61, pp. 38-45, 2015.

[3]. J. Han, M. Kamber, J. Pei, “*Data Mining: Concepts and Techniques*”, Morgan Kaufmann Publishers, 3rd Edition, India, 2011.

[4]. L.V. Bijuraj, "*Clustering and its Applications*", In the Proceedings of National Conference on New Horizons in IT – NCNHIT, India, pp. 169- 172, 2013.

[5]. P. Arora, Deepali, S.Varshney, "*Analysis of K-Means and K-Medoids Algorithm For Big Data*", International Conference on Information Security & Privacy, India, Science Direct, Vol. 78, pp. 507-512, 2016.

[6]. Z. Huang, "*Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*", ACM Transaction on Data Mining and Knowledge Discovery, Vol. 2, pp. 283–304, 1998.

[7]. Y. Sun, Q. Zhu, Z. Chen, "*An iterative initial-points refinement algorithm for categorical data clustering*", Pattern Recognition Letters, Elsevier, Vol. 23, Issue. 7, pp. 875–884, 2002.

[8]. D. Barbara, J. Coute, Yi Li, "*COOLCAT: An entropy based algorithm for categorical clustering*", Proceedings of the eleventh international conference on Information and knowledge management, USA, ACM, pp. 582-589, 2002.

[9]. O. M. San, V. Hyunh, Y. Nakamori, "*An Alternative Extension of the k-Means Algorithm for Clustering Categorical Data*". International Journal Applied Math and Computer Science, Vol.14, pp. 241–247, 2004.

[10]. F. Cao, J. Liang, L. Bai, "*A new initialization method for categorical data clustering*", Expert Systems with Applications, Science Direct, Vol. 36, pp. 10223-10228, 2009.

[11]. J. Lee, Y. J. Lee, M. Park, "*Clustering with Domain Value Dissimilarity for Categorical Data*", Advances in Data Mining, Applications and Theoretical Aspects, Lecture Notes in Computer Science, Springer, Vol. 5633, pp. 310-324, 2009.

[12]. D. Ienco, R. G. Pensa, R. Meo, "*From Context to Distance: Learning Dissimilarity for Categorical Data Clustering*", ACM Transactions on Knowledge Discovery from Data, pp.1-22, 2011.

[13]. A. Desai, H. Singh, V. Pudi, "*DISC: Data Intensive Similarity Measure for Categorical Data*", Proceedings of Advances in Knowledge Discovery and Data Mining – 15th Pacific Asia Conference, Springer, pp. 469 – 481, 2011.

[14]. F. Cao, J. Liang, D. Li, L. Bai, C. Dang, "*A dissimilarity measure for the k-modes clustering algorithm*", Knowledge-Based Systems, Elsevier, Vol. 26, pp. 120–127, 2012.

[15]. Y. M. Cheung, H. Jia, "*Categorical and numerical attribute data clustering based on a unified similarity metric without knowing cluster number*", Pattern Recognition, Elsevier, Vol. 46, pp. 2228–2238, 2013.

[16]. S. S. Khan, A. Ahmad, "*Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering*", Expert Systems with Applications, Elsevier, Vol. 40, pp. 7444–7456, 2013.

[17]. R. S. Sangam, H. Om, "*The k-modes algorithm with entropy based similarity coefficient*", 2nd International Symposium on Big Data and Cloud Computing, Procedia Computer Science, Elsevier, Vol. 50, pp. 93-98, 2015.

[18]. R.Viederyte, "*Preconditions evaluation in Maritime Clustering research*", 3rd Global Conference on Business, Economics, Management and Tourism, Rome, Italy, Elsevier, Vol. 39, pp. 365-372, 2016.