Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7676990

DESIGN OF A CLUSTERING ALGORITHM FOR LARGE IOT DATASET

Dr.P.S.Smitha
Associate Professor, Dept.of CSE, Velammal Engineering College, Chennai-66,
smitha.p@velammal.edu.in

Mrs.T.Subashini
Assistant Professor, Dept.of CSE, Velammal Engineering College-66, Chennai,
subashini.t@velammal.edu.in

Mr. R.Akhil Nair
Assistant Professor, Dept.of CSE, Velammal Engineering College, Chennai- 66,
akhilnair@velammal.edu.in

Mrs.C.Sruthi Nath
Assistant Professor, Dept.of CSE, Velammal, Engineering College, Chennai - 66.
sruthinath@velammal.edu.in

Abstract
Data mining refers to the preset procedures and algorithms used to extract these valuable patterns. The research aims to improve partition-based clustering algorithmswith advanced features of efficient data analysis and automatically generate an appropriate number of clusters.The efficiency of K-Means clustering is further challenged by real-world datasets with high dimensionality. As a result, the algorithm becomes too expensive to implement. With an increase in size comes a decrease in cluster quality. This study proposes a K-Modes algorithm-based technique for working efficiently with large dimension datasets.Improvements can be made to this approach by eliminating non-significant features from the clustering process, which reduces the dimensionality of the clusters created and improves their accuracy. However, this number can be used as an input depending on user requirements if it has a significance value greater than or equal to 60% of the maximum significance value in the proposed algorithm.

Keywords: Data Mining, K-Means, Clustering, IoT

1. Introduction

Using data mining, patterns and knowledge may be collected from the massive amounts of data collected by information systems, and these patterns can be used to make sound judgments. Organizations can make better decisions because of it. Data mining offers methods for efficiently processing massive amounts of data and presenting it in the desired format. Data mining is used in Data Mining Techniques refers to the preset procedures and algorithms used to extract these valuable patterns [1].

Clustering is used in market surveys to find customers with specific purchasing patterns. In addition to the above stated industries, there is a lot of work being done to apply these

techniques elsewhere.The literature contains several different clustering algorithms. Hierarchical clustering algorithms and Partitional clustering algorithms are the two main types of clustering algorithms. In order to detect clusters, hierarchical algorithms arrange data in a hierarchy. These methods, however, are not ideal for huge datasets because of this. The partition-based clustering algorithms, on the other hand, locate the clusters on their own [2] [3].

To overcome this obstacle, a great deal of effort has been expended. Other restrictions include needing to enter a known number of clusters as an input; in this case, it is K. When the dataset is fresh or unknown, an erroneous number of clusters is required, resulting in wasteful data grouping. Researchers are continuously looking for ways to get around this obstacle. When working with categorical data, K-Mean constraints are carried over to its K-Modes.High dimensional datasets also pose a significant barrier to K-Mode efficiency.

2. Related works

Researchers have been working on finding a suitable value for the value of K, the number of clusters formed. The value is often a wild guess from the user as he does not have much awareness about the dataset and its properties. The recently introduced K-means clustering is an attempt in estimating a suitable value. The K-means algorithm (Lakshmi & Raju, 2015) can be considered as a pre-processing step to predict the value of number of clusters to be formed on the dataset [4].

The algorithm is capable of handling ordinal type of data but has major difficulties in the case of continuous data. The aim of the study is to provide some more capabilities to the basic K-means algorithm to handle even the real valued data effectively in forming the clusters. The focus of the study is to ascertain the usefulness of Clustering in predicting a suitable value for K, the most important parameter of K-Means algorithm [5].

Computation of similarity between any two instances of a dataset can be performed based on the concept of similarity. Most of the time, the dataset to be clustered contain different attributes of different data types such as qualitative, categorical and nominal. If one wishes to use distance-based measure for calculation, there needs to be some pre-processing for the non-numeric data into numeric representation before the actual clustering starts [6].

For handling non-numeric as well as mixed mode data in datasets, and for measuring similarity between any two given instances in the dataset, the concept of Jaccard similarity index can be effectively utilized (The Jaccard similarity index can be defined as the size of intersection (the number of attributes with matching values) divided by the size of the union of the attribute sets[7].

For numerical data, especially with integer values, we can easily find out the tuples (instances of object) having same or exact value for attributes when compared with the base tuple selected from the dataset. Here one of the tuples from the dataset is selected as the base object and then find those subsets of tuples similar in value to form the cluster. For string type of data, finding similarity is bit more troublesome and one needs to use additional measures for checking their equality or similarity [8].

For better clustering, similarity needs to be calculated for datasets, with even a slight variation in attribute values getting reflected on clustering. The clustering method needs to compare attribute values of objects to see whether they are same or not and then calculate similarity index. This method has so many advantages over distance-based measures as it can be applied on any attribute, irrespective of its data type [9].

The K-means clustering algorithm uses the concept of Jaccards coefficient of similarity to find the similar instances based on attribute values of the given instance becoming equal to that of the chosen base vector. The algorithm is designed to check for an exact match in attribute values of all other instances and clustering is done based on the measure of similarity existing between attributes. The user specifies an upper bound for the number of clusters to be formed [10].

The clustering process uses this value to create equal intervals for cluster formation. The degree of similarity of all instances is estimated by comparing to the base vector and instances are put into appropriate matching cluster. There is every likelihood of forming lesser number of clusters than the specified value, as there may not be instances in some of the intervals of similarity [11].

A major limitation or drawback in the working of K-means clustering is that the last cluster, which contain all the dissimilar objects as compared to the base object, appears to be too much crowded. The last cluster is highly populated since the search is for an exact match in finding the similarity and as per the algorithm, all those tuples which non-matches are put into the last cluster. The last cluster contains all those objects which are non-similar to the base vector and the intra cluster similarity is too low. Ideally a clustering operation should result in clusters with high inter cluster similarity [12].

The clustering process is expected to generate clusters of high degree of cohesion, with maximum similarity existing between cluster members. The last cluster formed in K-means clustering contains all those instances dissimilar to the base vector. The elements in this cluster may not be having any noticeable similarity and the overall quality of clustering will be affected.

Another limitation to the algorithm is the choice of base vector. It is evident that the choice of base vector has a critical role in the formation of clusters. If the base vector chosen is a good representative of all other instances, getting good clusters is almost assured. If the chosen vector is an outlier, then the whole clustering process may suffer. The algorithm by default, chooses the first instance as the base vector and the chances of that becoming a good representative is very low. There is scope for potential research work in the selection of the base vector.

The author proposed novel numerical, categorical, and mixed dataset methods based on a family of algorithms known as the K-Means family that work without requiring any parameters to be supplied.The efficiency of K-Means clustering is put to the test by real-world datasets with high dimensionality. Increasing the cluster size increases computing costs and degrades the quality of the clusters. Even though this issue has been addressed for numerical datasets,

little progress has been made in dealing with categorical and mixed datasets. The author came up with a method based on the K-Modes technique that works well with large datasets.

This research aims to improve partition-based clustering algorithms, particularly K-Means, K-Modes and K-Prototype, with advanced features of efficient data analysis and automatically generate an appropriate number of clusters for numerical, categorical and mixed datasets.The efficiency of K-Means clustering is further challenged by real-world datasets with high dimensionality. As a result, the algorithm becomes too expensive to implement. With an increase in size comes a decrease in cluster quality. This study proposes a K-Modes algorithm-based technique for working efficiently with large dimension datasets.

3. Methods

Numerical datasets, categorical datasets, and mixed datasets all have a problem because traditional methods need you to provide the value of K as an input. A new algorithm is proposed that does not require K as an input and performs well with large categorical datasets. C# is used to implement the recommended algorithms.

Methods provided in this paper yield clusters with more accuracy than the original K-Means, K-Modes, and K-Prototype algorithms, as well as many extensions proposed by different authors.

A complete business analytics workbench, RapidMiner is an open source system with an emphasis on data mining, text analysis and predictive analytics. Many descriptive and predictive techniques are used to offer you the information you need. With RapidMiner and RapidAnalytics, a complete business intelligence solution with predictive analytics is available with full reporting and dash boarding capabilities. Real datasets from UCI Machine Repository: a website that provides the machine learning community with 300 datasets are used in the experiments.

New algorithms built on the partition-based clustering algorithm (K-Means) are presented, but with advanced features for efficient data analysis and automatic generation of the appropriate number of clusters for numerical and categorical and mixed datasets. Additionally, the newly developed algorithms produce clusters that are equivalent to those produced by the original methods and a variety of additional algorithms as well. The K-Modes approach is extended with a new algorithm that does not require K as an input and also works well with large datasets.

K-Means

Instead of starting with an estimate and specification of how many groups (K) should be generated as input, we propose using a K-Means approach that can yield significantly more meaningful clusters. This approach performs best with datasets that have all of their attributes measured on the same numeric scale.

Unlike the previous approaches, this one only requires the dataset itself as an input. In the beginning, the dataset is divided into two clusters by selecting two items as initial centroids in this approach. Each object total attribute value represents a range from 0 to 1 (inclusive). Based

on Euclidean distance, this objective function evaluates a set of inputs. The proposed algorithm pseudo code is detailed in the following sections.

Input: The set of n object attributes as an input

Output: Clusters of objects from the input dataset are returned as an output.

Step 1: Compute the total of each object attribute values.

Step 2: Use objects with values as initial centroids.

Step 3: Use the Euclidean distance between and the centroids to divide the data into two initial divisions.

Step 4: Compute the average of the Euclidean distances between each cluster objects and their centroid.

Step 5: In step 3, you constructed partitions by computing new means (centroids).

Step 6: In order to discover outliers, use the updated centroids to calculate the Euclidean distance between each item and them.

Step 7: Capture the reshaped clusters.

Step 8: Calculate the transformed cluster centroids.

Step 9: Find the outliers that do not meet the goal function in step 6 by calculating the Euclidean distance between each outlier and each of the cluster centroids.

Step 10: Restore the outliers that meet the objective function and you'll have a new set of clusters that are altered.

Step 11: This is the final phase. carried out to see whether any of the outliers in the current clusters may be altered. Set B now contains all of the remaining outliers.

As a result, the proposed method does not require an initial estimation or definition of the number of clusters. This algorithm begins by selecting two items as initial centroids and partitioning the dataset into two clusters.

According to experiments on real datasets of varying sizes and dimensions, the suggested K-Means method produces better clusters than basic K-Means and a variety of other algorithms, even when there is no prior estimate of K.

K-Modes Algorithm

Before running the procedure, the dataset is divided into a few clusters. Clusters are analysed for outliers, or objects that fail to meet the objective function. A criterion outlined later in this chapter is used to see if any of the outliers found can be returned to the clusters. When an object cannot meet the objective function due to the presence of a neutral object, this step is performed to allow the object to be placed back in the cluster. Once the outliers have been removed, a new dataset is created from which all objects can be assigned to a cluster that meets the objective function.

The proposed technique utilises Ahmad et al. (2007) idea of attribute significance to generate initial clusters. The centroids of the clusters are found using Huang (1998) frequency-based technique, and the distance between an object and the centroid is determined by San et al. using their distance measure (2004).

To deal with large datasets, the second version of the technique has a dimensionality reduction step to ensure that only relevant attributes (those with a high significance value) are used in the clustering process.

Pseudocode

Input: A dataset (D) containing N items with m categorical properties

Output: Clusters are used to distribute the objects.

Step 1: Create the first set of clusters.

Step 2: Find the cluster centroids using step 3.

Step 3: Find the distance between the centroid of each object and that distance.

Step 4: In order to find the outliers in the original clusters, use the objective function.

Step 5: Obtain clusters that have undergone transformation.

Step 6: Find the number of characteristics in each cluster where objects have the same value.

Step 7: The sum of these numbers is given as n. Only include outliers from the converted clusters in step 5 that have values matching in n-1 characteristics with the outliers from those clusters.

Step 8: Set B now includes all of the remaining outliers.

Step 9: Find the number of outliers in each cluster by repeating step 8 multiple times.

Step 10: If k is more than 1, then

a. As long as two outliers have identical values but are located at different distances from the centroid, keep them together in the cluster.

b. Set B should include a few more outliers.

c. If k is greater than 1, then nothing happens.

i. Until then, reintroduce the outlier into the cluster.

ii. The set of outliers

iii. Create a new cluster if set B contains a single object;

d. else,

i.keep the existing cluster.

e. End

Step 11: Set B is a new dataset that can be used in place of set a.

Step 12: Follow the steps 1 through 9 until B is reached.

4. COMPARATIVE ANALYSIS

There are several datasets used to test the method, including the Credit Approval and Zoo datasets. The suggested algorithm findings are compared to those of the original K-Modes, which were generated by RapidMiner. The known facts were used to determine the initial value of K.

Results on Credit Approval Dataset

This dataset findings were achieved using the original K-Modes, a proposed algorithm without dimensionality reduction, and a proposed technique that does so.

Table 1: Results for Credit Approval Dataset with Dimensionality Reduction Feature

	K-Modes algorithm with K=2 (9 attributes processed)		Proposed algorithm without dimensionality reduction (9 attributes processed)				Proposed algorithm with dimensionality reduction (4 attributes processed)			
Cluster No.	1	2	1	2	3	4	1	2	3	4
No. of matching attributes	2	2	2	3	6	5	2	3	2	3
No. of records per cluster	507	158	498	156	7	4	207	300	126	32

Dimensionality Reduction Feature K-Modes Algorithm Results for the Credit Approval Dataset with K=2

Table 1 shows that the K-Modes algorithm clusters contain at most two objects matching in their attributes. To see if the results are comparable, we can look at the proposed technique that

includes a dimensionality reduction step. There are four clusters formed by the proposed algorithm, and cluster3 has records matching in six out of nine attributes, representing the dataset most comparable records. It was found that the original K-Modes and the proposed technique did not reduce the dimensions. Therefore, nine characteristics were included in clustering; however, with dimensionality reduction, only four of the most significant attributes were involved.

Table 2: Comparison of Clustering Accuracy for Credit Approval Dataset

Clustering Algorithm	Clustering Accuracy (%)
Hard K-Modes	75.17
Weighting K-Modes	75.13
Fuzzy K-Modes	74.91
New Fuzzy K – Modes	77.01
Liao et al. method	79.3
Bai et al. Method	78.41
Proposed algorithm without dimensionality reduction	71.27
Proposed algorithm with dimensionality reduction	82.55

Table 2shows that the accuracy has improved for high-dimension datasets with this version of the proposed approach.

Results on Zoo Dataset

Attributes are listed from 1 to 16, and objects are listed from 1 to 101. There are seven preset classes, but 60 of the 101 instances are split into two groups. Table 3summarises the findings. RapidMiner initial K-Modes algorithm builds clusters with matching values for two attributes when K = 2. All objects were grouped together with no objects matching in any of the properties when K>2.

Five clusters are generated using the proposed technique, which does not include a dimensionality reduction step. Five clusters are also generated by the proposed algorithm dimensionality reduction stage. In addition, the technique with dimensionality reduction uses 11 of the total 16 attributes for clustering.

Table 3: Results for Zoo Dataset

	K-Modes algorithm (16 attributes processed)		Proposed algorithm without dimensionality reduction (16 attributes processed)					Proposed algorithm with dimensionality reduction (11 attributes processed)				
Cluster No.	1	2	1	2	3	4	5	1	2	3	4	5
No. of matching attributes	2	2	2	5	15	15	-	3	4	9	6	11
No. of records per cluster	73	28	67	28	3	2	1	36	41	15	5	4

As shown in Table 3, the proposed algorithm clusters are more accurate than those created by several existing techniques.

5. Conclusion

Using the algorithm discussed in this chapter, a suitable number of clusters can be generated without having to know in advance how many clusters there are. In terms of cluster quality, the new results outperform the previous K-Modes. Clusters have been formed by the algorithm so that items inside a cluster have the greatest possible similarity in terms of the number of characteristics that share the same value. It is possible to generate a cluster of objects with maximum similarity from a dataset without specifying the initial value of K in this manner. Additional improvements can be made to this approach by eliminating non-significant features from the clustering process, which reduces the dimensionality of the clusters created and improves their accuracy. Those qualities in the suggested method that have a significance greater than or equal to 60% of the maximum significance value are taken into account. This number can be used as an input if needed.

REFERENCES

[1] Abubaker, Mohamed, and Ashour, Wesam 2013, Efficient Data Clustering Algorithms: Improvements over Kmeans', International Journal of Intelligent Systems and Applications, 5, 3, pp. 37-49.

[2] Bai, Liang, Liang, Jiye, Dang, Chuangyin and Cao, Fuyuan 2013, A novel fuzzy clustering algorithm with between-cluster information for categorical data', Fuzzy Sets and Systems, 215, pp.55-73.

[3] Cheung, Y. and Jia, H. 2013, Categorical-and-Numerical-Attribute Data Clustering based on a Unified Similarity Metric without Knowing Cluster Number', Pattern Recognition, 46, pp. 2228–2238.

[4] Han, Jiawei, Kamber, Micheline and Pei, Jian 2011, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[5] Ienco, Dino, Pensa, G., Riggero and Meo, Rosa 2012, From Context to Distance: Learning Dissimilarity for Categorical Data Clustering', ACM Transactions on Knowledge Discovery from Data, 6, 1, pp. 1-25.

[6] Kaur, Harleen and Wasan, Krishan, Siri 2006, Empirical Study on Applications of Data Mining Techniques in Healthcare', Journal of Computer Science, 2, 2, pp. 194- 200.

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

[8] Liang, Jiye, Zhao, Xingwang, Li, Deyu, Cao, Fuyuan and Dang, Chuangyin 2012, Determining the Number of Clusters using Information Entropy for Mixed Data', Pattern Recognition, 45, 6, pp.2251–2265.

[9] Ma, Yiming, Liu, Bing, Wong, Kian, Ching, Yu, S., Philip and Lee, Ming, Shuik 2000, Targeting the right students using data mining', Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 457-464.

[10] Naija, Yosr, Chakhar, Salem, Blibech, Kaouther and Robbana, Riadh 2008, Extension of Partitional Clustering Methods for Handling Mixed Data', IEEE International Conference on Data Mining Workshops, pp. 257-266.

[11] Ogor N. Emmanuel 2007, Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques', Fourth Congress of Electronics, Robotics and Automotive Mechanics, pp. 354-359.

[12] Panda, Sandeep, Sahu, Sanat, Jena, Pradeep and Chattopadhyay, Subhagata 2012, Comparing Fuzzy-C Means and K-Means Clustering Techniques: A Comprehensive Study', Advances in Intelligent and Soft Computing, 166, pp. 451-460.