Project Report

On

**CLUSTERING OF MIXED DATA USING ENTROPY BASED**

**K-MODE ALGORITHM**

Submitted as a part of Major Project

In fulfillment of the requirement for the award of the degree of

Bachelor of Technology-Information Technology [IT]



*Submitted to[Mentor]: Submitted by:*

**Ms. Juhi Singh Karan Chopra (05710403114)**

**Sagar Jha (06210403114)**

**Vivek Khandelwal (05610403114)**

**Mayank Tagra (06310403114)**

Department of IT and CS

**AMITY SCHOOL OF ENGINEERING AND TECHNOLOGY**

(Affiliated to Guru Gobind Singh Indraprastha University)

(Jan-May, 2018)

**CERTIFICATE**

This is to certify that below listed students of B.Tech 4th year from Information Technology have successfully completed their Major Project entitled “**Clustering Of Mixed Data Using Entropy Based K-Mode Algorithm** “.They have submitted their Project Report for the partial fulfillment of the curriculum of the Degree of Bachelor of Technology from Amity School Of Engineering &Technology.

**S.NO NAME ENROLLMENT NUMBER**

1 Karan Chopra 05710403114

2 Sagar Jha 06210403114

3 Vivek Khandelwal 05610403114

4 Mayank Tagra 06310403114

*Ms. Juhi Singh Prof. M.N. Gupta*

*(ASST. Professor) (Head of Department of IT and CS)*

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of the project would be incomplete without mentioning the names of the people who made it possible and whose constant guidance and encouragement crowns all efforts with our success.

We extend our gratitude to **Prof. (Dr.) Rekha Aggarwal** , Director and **Prof. M.N. Gupta**, Head of Department of IT and CS , Amity School of Engineering and Technology for providing us with excellent infrastructure and constructive environment which laid a potentially strong foundation of our professional life.

We would like to express our profound thanks to **Ms. Juhi Singh** who guided us throughout the project like an expert caption on a ship providing us each and every detail, reference and technical help, without which it was difficult to accomplish.

*Karan Chopra Sagar Jha Vivek Khandelwal Mayank Tagra*

(05710403114) (06210403114) (05610403114) (06310403114)

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **FIGURE** | **NAME** | **PAGE NUMBER** |
| 1.1  1.2  1.3  1.4  1.5  3.1  5.1  6.1 | Data mining  Stages of Data Mining  Data Mining Issues  Applications of Data Mining  Clustering  The result of k-means  Mixed data variables  Data set | 7  8  12  12  13  22  29  32 |

**TABLE OF CONTENTS**

Certificate------------------------------------------------------------------------------------------ 2

Acknowledgement-------------------------------------------------------------------------------- 3

List of figures ------------------------------------------------------------------------------------- 4

Abstract--------------------------------------------------------------------------------------------- 6

|  |  |  |
| --- | --- | --- |
| **Chapter** | **Topic** | **Page Number** |
| **1**  **2**  **3**  **4**  **5**  **6**    **7** | **Introduction**  1.1 Data Mining  1.1.1 Data Mining Parameters  1.1.2 Data Mining Tasks  1.1.3 Data Mining Issues  1.1.4 Applications of Data Mining  1.2 Cluster Analysis  1.2.1 What is Clustering?  1.2.2 Clustering Methods  **Literature Survey**  2.1 Related Works  **Existing Algorithms**  3.1 K-Means  3.1.1 Algorithm for K-Means Clustering  3.2 K-Mode  3.2.1 Algorithm for K-Means Clustering  **Proposed Algorithm**  4.1 EBK-Modes: Entropy Based K-Modes  4.2 Numerical Example on EBK-Mode  4.3 EBK-Mode Algorithm[categorical data]  **Entropy based k-mode [EBK] for mixed**  **Data**  5.1 Mixed data  5.2 Algorithm of EBK for Mixed data.  5.3 Flowchart  **Implementations**  6.1 Data Set  6.2 K-Means Implementation  6.3 K-Mode Implementation  6.4 EBK-Mode Implementation  6.5 EBK-mode implementation for mixed data  **Conclusions**  **REFERENCES** | **7**  7  8  9  11  12  13  13  15  **18**  18  **20**  20  21  23  23  **25**  25  25  28  **29**  29  29  31  **32**  32  33  35  36  38  **41**  **42** |

**ABSTRACT**

The interest in attribute weighting for soft subspace clustering have been increasing in the last years. Most of the proposed approaches are designed for dealing only with numeric data. In this our focus is on soft subspace clustering for categorical data. In soft subspace clustering, the attribute weighting approach plays a crucial role. We will implement an entropy-based approach for measuring the relevance of each categorical attribute in each cluster and will compare the performance of proposed algorithm.

Most of the recent results in soft subspace clustering for categorical data, proposes modification of k-modes algorithm. In general in these approaches the contribution of each attribute is measured considering only the frequency of the mode category of the average distance of the data objects from mode of cluster.

In the proposed work the strategy for measuring the contribution of each attribute considering the motion of entropy,which measures the uncertainity of a given random variable is explored.

We will implement the EBK-mode an extension of the basic knowledge algorithm that uses the notion of entropy for measuring the relevance of each attribute in each cluster.

After completing the work for the categorical data ,same will done for the mixed data .

The given theory will be implemented using java jdk1.8

**CHAPTER 1 INTRODUCTION**

**1.1 DATA MINING**

There is a huge amount of data available in the Information Industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it.

Extraction of information is not the only process we need to perform; data mining also involves other processes such as Data Cleaning, Data Integration, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation. Once all these processes are over, we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc.

Data mining is the process of sorting through large [data sets](http://whatis.techtarget.com/definition/data-set) to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. **Data mining** is the computing process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use

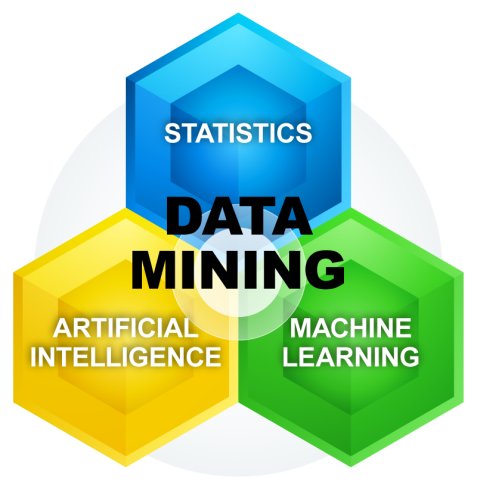


Fig 1.1 Data mining

**1.1.1 DATA MINING PARAMETERS**

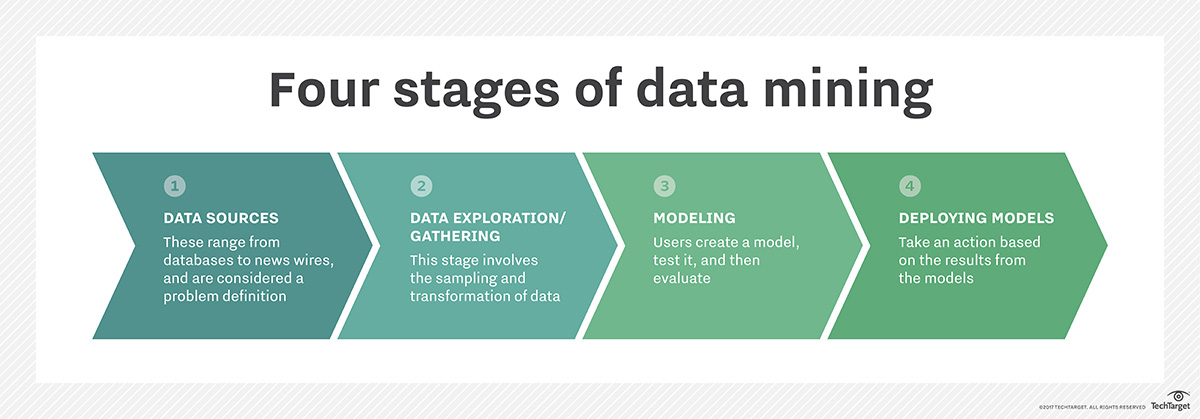
Data mining parameters include Sequence or Path Analysis, Classification, Clustering and Forecasting.

Sequence or Path Analysis parameters look for patterns where one event leads to another later event. A Sequence is an ordered list of sets of items, and it is a common type of data structure found in many databases.

A Classification parameter looks for new patterns, and might result in a change in the way the data is organized. Classification algorithms predict variables based on other factors within the database.

Clustering parameters find and visually document groups of facts that were previously unknown. Clustering groups a set of objects and aggregates them based on how similar they are to each other. There are different ways a user can implement the cluster, which differentiate between each clustering model.

Fostering parameters within data mining can discover patterns in data that can lead to reasonable predictions about the future, also known as predictive analysis

Fig 1.2 Stages of Data Mining

**1.1.2 DATA MINING TASKS**

Data mining deals with the kind of patterns that can be mined. On the basis of the kind of data to be mined, there are two categories of functions involved in Data Mining –

* Descriptive
* Classification and Prediction

**DESCRIPTIVE FUNCTIONS**

The descriptive function deals with the general properties of data in the database. Here is the list of descriptive functions −

* Class/Concept Description
* Mining of Frequent Patterns
* Mining of Associations
* Mining of Correlations
* Mining of Clusters

### Class/Concept Description

Class/Concept refers to the data to be associated with the classes or concepts. For example, in a company, the classes of items for sales include computer and printers, and concepts of customers include big spenders and budget spenders. Such descriptions of a class or a concept are called class/concept descriptions. These descriptions can be derived by the following two ways −

* **Data Characterization** − This refers to summarizing data of class under study. This class under study is called as Target Class.
* **Data Discrimination** − It refers to the mapping or classification of a class with some predefined group or class.

### Mining of Frequent Patterns

Frequent patterns are those patterns that occur frequently in transactional data. Here is the list of kind of frequent patterns −

* **Frequent Item Set** − It refers to a set of items that frequently appear together, for example, milk and bread.
* **Frequent Subsequence** − A sequence of patterns that occur frequently such as purchasing a camera is followed by memory card.
* **Frequent Sub Structure** − Substructure refers to different structural forms, such as graphs, trees, or lattices, which may be combined with item-sets or subsequences.

### Mining of Association

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

For example, a retailer generates an association rule that shows that 70% of time milk is sold with bread and only 30% of times biscuits are sold with bread.

### Mining of Correlations

It is a kind of additional analysis performed to uncover interesting statistical correlations between associated-attribute-value pairs or between two item sets to analyze that if they have positive, negative or no effect on each other.

### Mining of Clusters

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.

## Classification and Prediction

Classification is the process of finding a model that describes the data classes or concepts. The purpose is to be able to use this model to predict the class of objects whose class label is unknown. This derived model is based on the analysis of sets of training data. The derived model can be presented in the following forms −

* Classification (IF-THEN) Rules
* Decision Trees
* Mathematical Formulae
* Neural Networks

The list of functions involved in these processes are as follows −

* **Classification** − It predicts the class of objects whose class label is unknown. Its objective is to find a derived model that describes and distinguishes data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.
* **Prediction** − It is used to predict missing or unavailable numerical data values rather than class labels. Regression Analysis is generally used for prediction. Prediction can also be used for identification of distribution trends based on available data.
* **Outlier Analysis** − Outliers may be defined as the data objects that do not comply with the general behavior or model of the data available.
* **Evolution Analysis** − Evolution analysis refers to the description and model regularities or trends for objects whose behavior changes over time.

## Data Mining Task Primitives

* We can specify a data mining task in the form of a data mining query.
* This query is input to the system.
* A data mining query is defined in terms of data mining task primitives.

**1.1.3 DATA MINING ISSUES**

Data mining is not an easy task, as the algorithms used can get very complex and data is not always available at one place. It needs to be integrated from various heterogeneous data sources. These factors also create some issues. Here in this tutorial, we will discuss the major issues regarding −

* Mining Methodology and User Interaction
* Performance Issues
* Diverse Data Types Issues

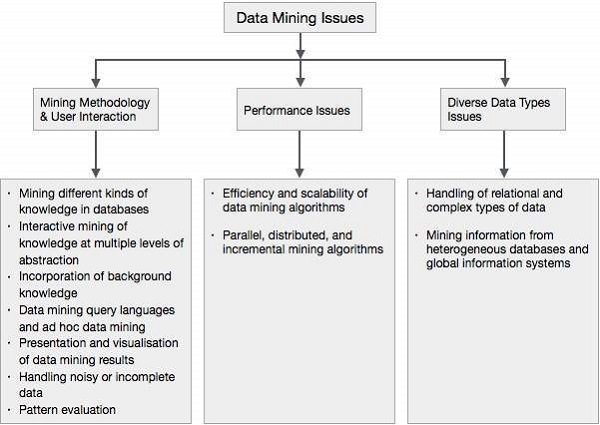


Fig 1.3 Data Mining Issues

**1.1.4 Applications of Data Mining**

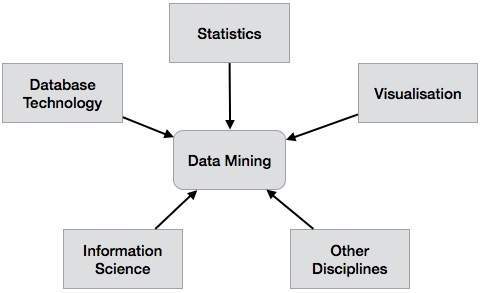


Fig 1.4 Applications of Data Mining

**1.2 CLUSTER ANALYSIS**

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.

## 1.2.1 What is Clustering?

Clustering is the process of making a group of abstract objects into classes of similar objects.

**Points to Remember**

* A cluster of data objects can be treated as one group.
* While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.
* The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

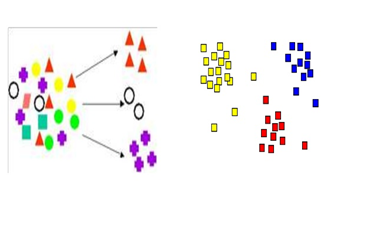


Fig 1.5 Clustering

## Applications of Cluster Analysis

* Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
* Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
* In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
* Clustering also helps in identification of areas of similar land use in an earth observation database. It also helps in the identification of groups of houses in a city according to house type, value, and geographic location.
* Clustering also helps in classifying documents on the web for information discovery.
* Clustering is also used in outlier detection applications such as detection of credit card fraud.
* As a data mining function, cluster analysis serves as a tool to gain insight into the distribution of data to observe characteristics of each cluster.

## Requirements of Clustering in Data Mining

The following points throw light on why clustering is required in data mining −

* **Scalability** −We need highly scalable clustering algorithms to deal with large databases.
* **Ability to deal with different kinds of attributes** − Algorithms should be capable to be applied on any kind of data such as interval-based (numerical) data, categorical, and binary data.
* **Discovery of clusters with attribute shape** −The clustering algorithm should be capable of detecting clusters of arbitrary shape. They should not be bounded to only distance measures that tend to find spherical cluster of small sizes.
* **High dimensionality** −The clustering algorithm should not only be able to handle low-dimensional data but also the high dimensional space.
* **Ability to deal with noisy data** − Databases contain noisy, missing or erroneous data. Some algorithms are sensitive to such data and may lead to poor quality clusters.
* **Interpretability** −The clustering results should be interpretable, comprehensible, and usable.

## 1.2.2 Clustering Methods

Clustering methods can be classified into the following categories −

* Partitioning Method
* Hierarchical Method
* Density-based Method
* Grid-Based Method
* Model-Based Method
* Constraint-based Method

### Partitioning Method

Suppose we are given a database of ‘n’ objects and the partitioning method constructs ‘k’ partition of data. Each partition will represent a cluster and k ≤ n. It means that it will classify the data into k groups, which satisfy the following requirements −

* Each group contains at least one object.
* Each object must belong to exactly one group.

**Points to remember −**

* For a given number of partitions (say k), the partitioning method will create an initial partitioning.
* Then it uses the iterative relocation technique to improve the partitioning by moving objects from one group to other.

### Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

### Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

### Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

### Density-based Method

This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

### Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.

**Advantage**

* The major advantage of this method is fast processing time.
* It is dependent only on the number of cells in each dimension in the quantized space.

### Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

### Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

**CHAPTER 2 LITERATURE SURVEY**

**2.1 RELATED WORKS**

In subspace clustering, objects are grouped into clusters according to subsets of dimensions (or attributes) of a data set [9]. These approaches involve two mains tasks, identiﬁcation of the subsets of dimensions where clusters can be found and discovery of the clusters from different subsets of dimensions. According to the ways with which the subsets of dimensions are identiﬁed, we can divide subspace clustering methods into two categories: hard subspace clustering and soft subspace clustering. The approaches of hard subspace clustering determine the exact subsets of attributes where clusters are discovered. On the other hand, approaches of soft subspace clustering determine the subsets of dimensions according to the contributions of the attributes in discovering the corresponding clusters. The contribution of a dimension is measured by a weight that is assigned to the dimension in the clustering process. The algorithm proposed in this paper can be viewed as a soft subspace clustering approach. In [12], for example, it is proposed an approach in which each weight is computed according to the average distance of data objects from the mode of a cluster. That is, it is assigned a larger weight to an attribute that has a smaller sum of the within cluster distances and a smaller weight to an attribute that has a larger sum of the within cluster distances. An analysis carried out by [4] have shown that this approach is sensitive to the setting of the parameter β. In [4], it is assumed that the weight of a given attribute for a given cluster is a function of the frequency of the categorical value of the mode of the cluster for that attribute. This approach requires the setting of three parameters (β,Tv and Ts) for determining the attribute weights. In [11], the authors use the notion of complement entropy for weighting the attributes. The complement entropy reﬂects the uncertainty of an object set with respect to an attribute (or attribute set), in a way that the bigger the complement entropy value is, the higher the uncertainty is. In [9] the authors noticed that the decrease of the entropy in a cluster implies the increase of certainty of a subset of dimensions with larger weights in determination of the cluster. According to this, their approach simultaneously minimize the within cluster dispersion and maximize the

negative weight entropy to stimulate more dimensions to contribute to the identiﬁcation of a cluster. In our approach, as in [9], we also use the notion of entropy for measuring the relevance of each attribute. However, here we assume that the relevance of a given attribute, for a given cluster, is inversely proportional to the average of the entropy that is induced by each attribute value of the mode of the cluster.

**CHAPTER 3 EXISTING ALGORITHMS**

There are two algorithms which were proposed before entropy based k-mode algorithm that were used for clustering of categorical data.

**3.1 K-MEANS**

**K-MEANS CLUSTERING**

* *K*-means clustering is used when you have unlabeled data (i.e., data without defined categories or groups).
* The goal of this algorithm is to find groups in the data, with the number of groups represented by the variable *K*.
* The algorithm works iteratively to assign each data point to one of *K* groups based on the features that are provided.

**Advantages:-**

* Fast, robust and easier to understand.
* Gives best result when data set are distinct or well separated from each other.
* If variables are huge, then  K-Means most of the times computationally faster.
* Holds true results even if some assumptions are false.

**Disadvantages:-**

* The learning algorithm requires apriori specification of the number of cluster centers.
* The use of Exclusive Assignment - If  there are two highly overlapping data then k-means will not be able to resolve       that there are two clusters.
* The learning algorithm is not invariant to non-linear transformationsi.e.with different representation of data we get different results (data represented in form of cartesian co-ordinates and polar co-ordinates will give different results).
* Euclidean distance measures can unequally weight underlying factors.
* The learning algorithm provides the local optima of the squared error function.
* Randomly choosing of the cluster center cannot lead us to the fruitful result.
* Applicable only when mean is defined i.e. fails for categorical data.
* Unable to handle noisy data and outliers*.*
* Algorithm fails for non-linear data set.

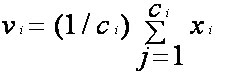
**3.1.1 ALGORITHM FOR K-MEANS CLUSTERING**

Let  X = {x1,x2,x3,……..,xn} be the set of data points and V = {v1,v2,…….,vc} be the set of centers.

1) Randomly select ‘c’ cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.

4) Recalculate the new cluster center using:    

Where *‘ci’* represents the number of data points in *ith* cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

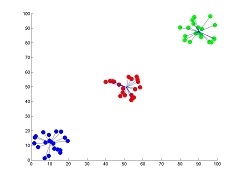


Fig 3.1 the result of k-means for *'N'* = 60 and *'c'* = 3

**K-MEANS WORKING**

k-means is  one of  the simplest unsupervised  learning  algorithms  that  solve  the well  known clustering problem. The procedure follows a simple and easy  way  to classify a given data set  through a certain number of  clusters (assume k clusters) fixed apriori. The main  idea  is to define k centers, one for each cluster. These centers should be placed in a cunning  way  because of  different  location  causes different  result. So, the better  choice  is  to place them  as  much as possible  far away from each other. The next step is to take each point belonging  to a  given data set and associate it to the nearest center. When no point is  pending,  the first step is completed and an early group age  is done. At this point we need to re-calculate k new centroids as barycenter of  the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points  and  the nearest new center. A loop has been generated. As a result of  this loop we  may  notice that the k centers change their location step by step until no more changes  are done or  in  other words centers do not move any more. Finally, this  algorithm  aims at  minimizing  an objective function know as squared error function.

**3.2 K-MODE**

The k-modes clustering algorithm is an extension to the standard k-means clustering algorithm for clustering categorical data. In data mining, k-means is the mostly used algorithm for clustering data because of its efficiency in clustering very large data. However, the standard k-means clustering process cannot be applied to categorical data due to the Euclidean distance function and use of means to represent cluster centres. To use k-means to cluster categorical data ,we convert each unique category to a dummy binary attribute and used 0 or 1 to indicate the categorical value either absent or present in a data record. This approach is not suitable for high dimensional categorical data. The k-modes approach modifies the standard k-means process for clustering categorical data by replacing the Euclidean distance function with the simple matching dissimilarity measure, using modes to represent cluster centres and updating modes with the most frequent categorical values in each of iterations of the clustering process. These modifications guarantee that the clustering process converges to a local minimal result and the efficiency of the clustering process is maintained.

Important points-

* Extension of K-means
* It works on numeric as well as non numeric values.
* Frequency based methods are used for clustering of objects.
* Replaces the means of clusters with modes.
* Uses simple matching dissimilarity measure for categorial data**.**

**3.2.1 ALGORITHM FOR K-MODE CLUSTERING**

To cluster a categorical data set X into k clusters, the k-modes clustering process consists of the following steps:

Step 1: Randomly select k unique objects as the initial cluster centres (modes).

Step 2: Calculate the distances between each object and the cluster mode; assign the object to the cluster whose centre has the shortest distance to the object; repeat this step until all objects are assigned to clusters.

Step 3: Select a new mode for each cluster and compare it with the previous mode. If different, go back to Step 2; otherwise, stop. This clustering process minimises the following k-modes objective function

U = [u ] i, j is an( n×k ) partition matrix, Z = {Z1 , Z2 , … , Zk } is a set of mode vectors and the distance function D(.,.) is defined as either (2) or (3).

Since it is essentially same as k-means, the k-modes clustering algorithm is efficient in clustering large categorical data and also produces locally minimal clustering result

**K-MODE WORKING**

The data is clustered by the k-mode method which aims to split the items into n groups such that the distance between the item and the the assigned cluster modes is minimised. Conventionally, simple matching distance is used for determination of the dissimilarity between two items or clusters. The computation takes place by keeping a count of the number of mismatches in the variables encountered. As an alternative, the weighted version of this distance is calculated by the frequencies of the categorical data . If a preset matrix is provided, it might be possible that none of the objects will be closest to one or more modes. In this case few clusters than supplied modes will be returned, intimating a warning.

**CHAPTER 4 PROPOSED ALGORITHM**

**4.1 EBK-MODES: ENTROPY BASED K-MODES.**

* Soft subspace clustering algorithm-For handling the high-dimensionality, some works take advantage of the fact that clusters usually occur in a subspace defined by a subset of the initially selected attributes.
* Extends the basic k-modes algorithm.
* Measures the contribution of each attribute using the notion of entropy.In Information Theory, entropy is a measure of the uncertainty in a random variable.The larger the entropy of a given random variable, the larger is the uncertainty associated to it.

**ENTROPY**

Considering a finite sample, the entropy H of a given variable X can be written as:

**H(X)=-∑P(Xi)log(P(Xi))**

where P(Xi) is the probability mass function

**4.2 NUMERICAL EXAMPLE ON EBK MODE**

Let us consider the following cluster c1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DATA OBJECT** | **A1** | **A2** | **A3** | **A4** | **A5** |
| **X1** | a | k | n | q | s |
| **X2** | b | k | n | r | s |
| **X3** | c | k | n | q | s |
| **X4** | d | k | o | r | t |

Its center is given by the mode of each attribute**: z1=(d,k,n,r,s)**

**APPROACH FOR ATTRIBUTE WEIGHTING**

* Let us consider ah(l) as the most frequently occurring value (the mode) of the attribute l, for a cluster ci.
* This is represented by the function φi(zij).

Maps a given categorical value ,which contains every object in the partition ci ∈ C.

For example φ1(s) = {x1,x2,x3}

The set of objects that have the value s (for the attribute a5).

**(ah(l),aj(p))**

A function that maps two given categorical values ah(l)∈ dom(ah) and aj(p)∈ dom(aj) to the number of objects, in ci ∈ C

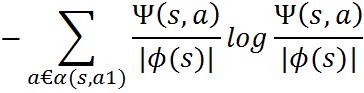
For example : ψ1(s,q)=|{x1,x3}| =2 .

**α(ah(l),aj)**

Maps a given categorical value and a given attribute

For example : α1(s,a1)={a,b,c}

**We can measure the entropy of any attribute ai, considering the set induced by φi(zij).  
 This is measured through a function Ei(zij,ai).**



For example-

ε1(s,a1)= −𝟏/𝟑𝒍𝒐𝒈𝟏/𝟑+𝟏/𝟑𝒍𝒐𝒈𝟏/𝟑+𝟏/𝟑𝒍𝒐𝒈𝟏/𝟑=𝟏.𝟏𝟎

ε1(s,a3)= −𝟑/𝟑𝒍𝒐𝒈𝟑/𝟑=𝟎

**The average of the uncertainty that is projected to a given attribute ah, considering the modes of all attributes, in a partition ci.**

* This is measured through a function:



where |A| is the number of attributes

FOR EXAMPLE:

E1(d,a5) = 0

E1(k,a5) = 0.56

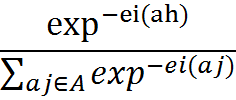
E1(n,a5) = 0 E1(a5)= [0+0.56+0+0.69+0]/5= 0.25

E1(r,a5) = 0.69

E1(s,a5) = 0

**ENTROPY-BASED RELEVANCE INDEX (ERI)**

* **It is inversely proportional to the E(ah), i.e., the average of the uncertainty that is projected to a given attribute ah, considering the modes of all attributes, in a partition ci.**

****

**ERIi(ah)=**

**DISSIMILARITY FUNCTION**

* EBK-modes adopts a function d that computes the dissimilarity between an object **xi** and a cluster mode **zl**.  **𝑑(xi, zl)= ∑ 𝛳 𝑎𝑗(xi, zl )**

Where  **𝜃(xi, zl )= 1 , Xij ≠ Zij**

**1−𝐸𝑅𝐼𝑎𝑗 , Xij = Zij**

**4.3 EBK-MODE ALGORITHM [categorical data]**

* **INPUT**: We have a set of data objects and number of clusters.
* Firstly we will take an empty k\*|A| array named as OldNodes.
* Then we will randomly choose K distinct objects from given set of data objects and assign them to the NewNodes.
* Till the point OldNodes≠NewNodes . For Every data object and for every cluster we will calculate dissimilarity between the data object and the mode and classify the data object into the cluster whose mode is nearest to it.

For every Cluster

* Calculate the mode of each cluster and assign to the NewNodes

Calculate weight of each attribute ah  ∈ A of the Cluster using ERI(ah ).

* **OUTPUT**: All the data objects are partitioned into k Clusters.

**CHAPTER 5 ENTROPY BASED K-MODE [EBK] FOR MIXED DATA**

**5.1 Mixed data**

* Mixed data – It is a type of data which has both numerical as well as categorical data like string.
* EBK[mixed data]= EBK[numerical data] + EBK[categorical data]
* Example of mixed data**:-**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **A** | **8** | **J** | **0** | **G** | **5** |

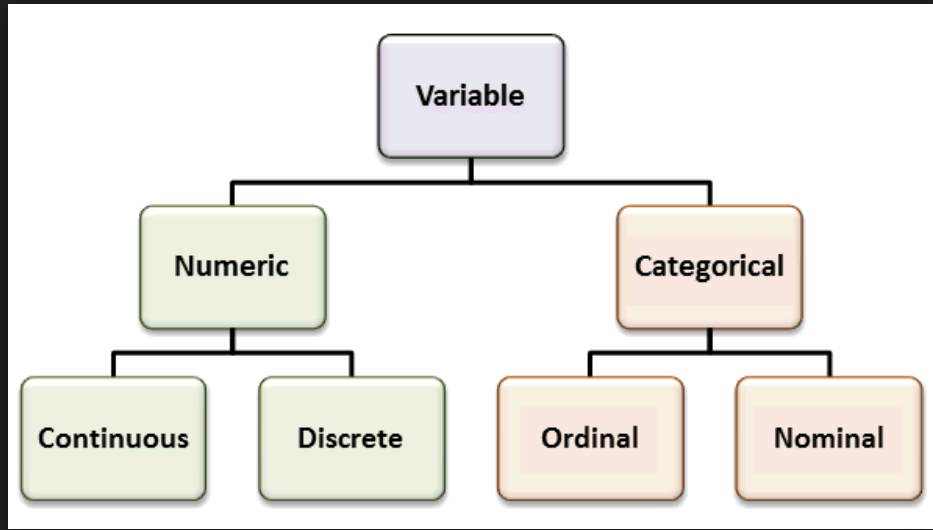
****

Fig 5.1 Mixed data variables

**5.2 Algorithm of EBK for Mixed data.**

* First we will declare object group and items.
* if(next element is a string)

{

if(string of object group)

{ store value in object group 1}

else

{ store value in item 1}

for each( row of object group 1)

{ execute with item 1 of next row of object group 1}

Calculate distance using dissimilarity function.

Display result 1

}

else

{

* if(object group numerical value)

{ store values in object group 2}

else { store values in item 2}

for each (row of object group 2)

{execute with item 1 of next row of object group 1}

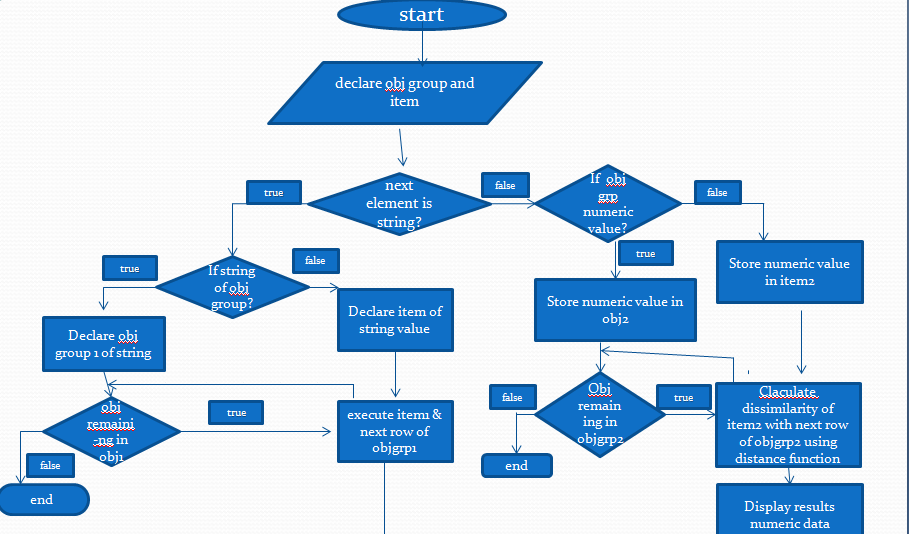
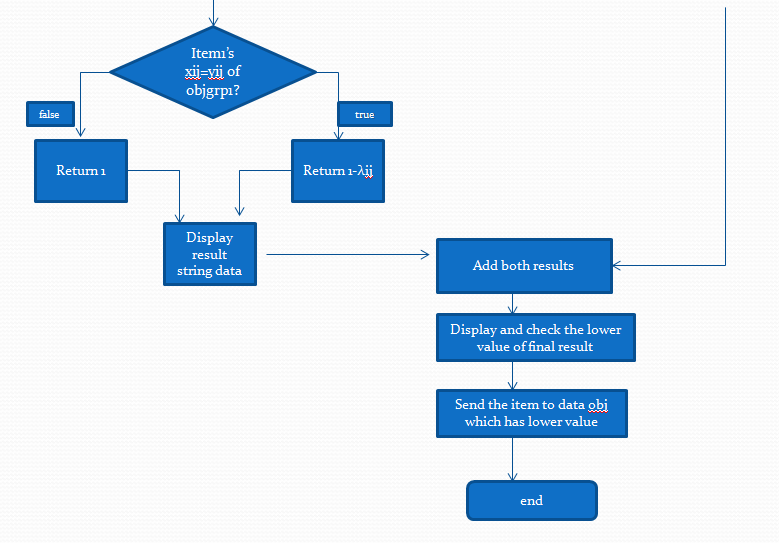
Calculate distance using distance function

Display result 2

}

* Add result 1 and result 2
* Store it in final result
* Select minimum of final result
* Put item in object group

**5.3 Flowchart**

****  

**CHAPTER 6 IMPLEMENTATIONS**

**6.1 DATA SET**

We used a data set named as the Chess data set.

This data set has been taken from the UCI-Machine learning repository[7]

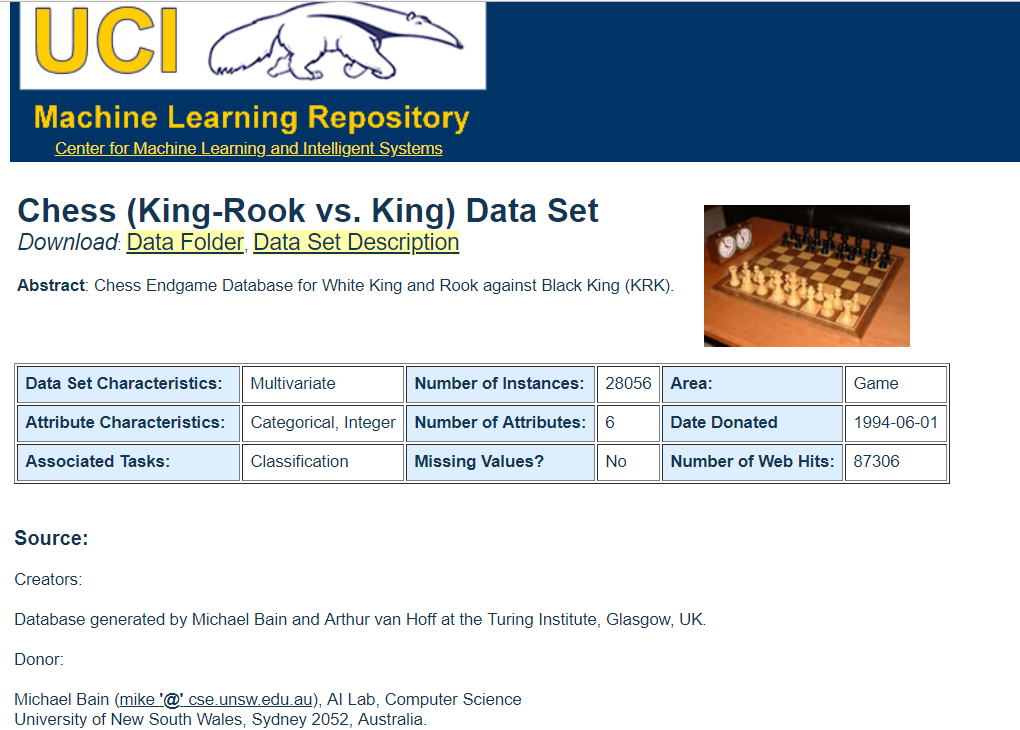
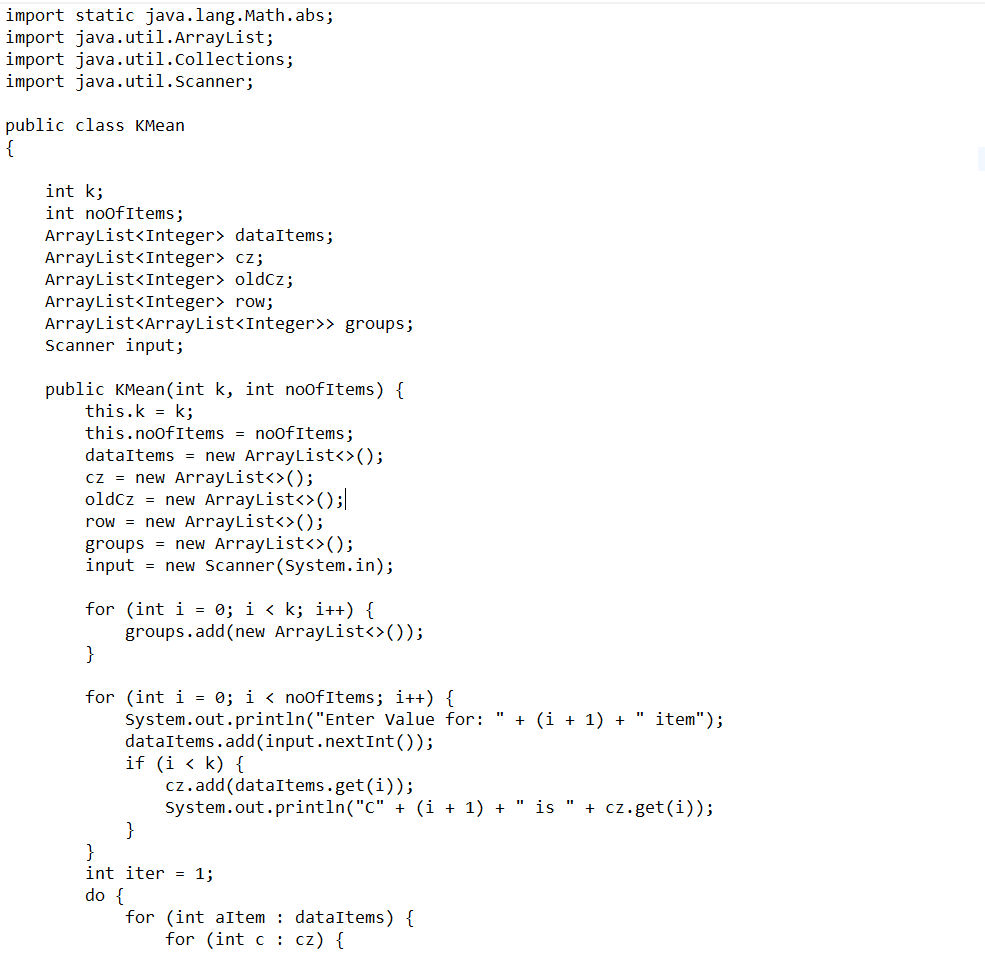
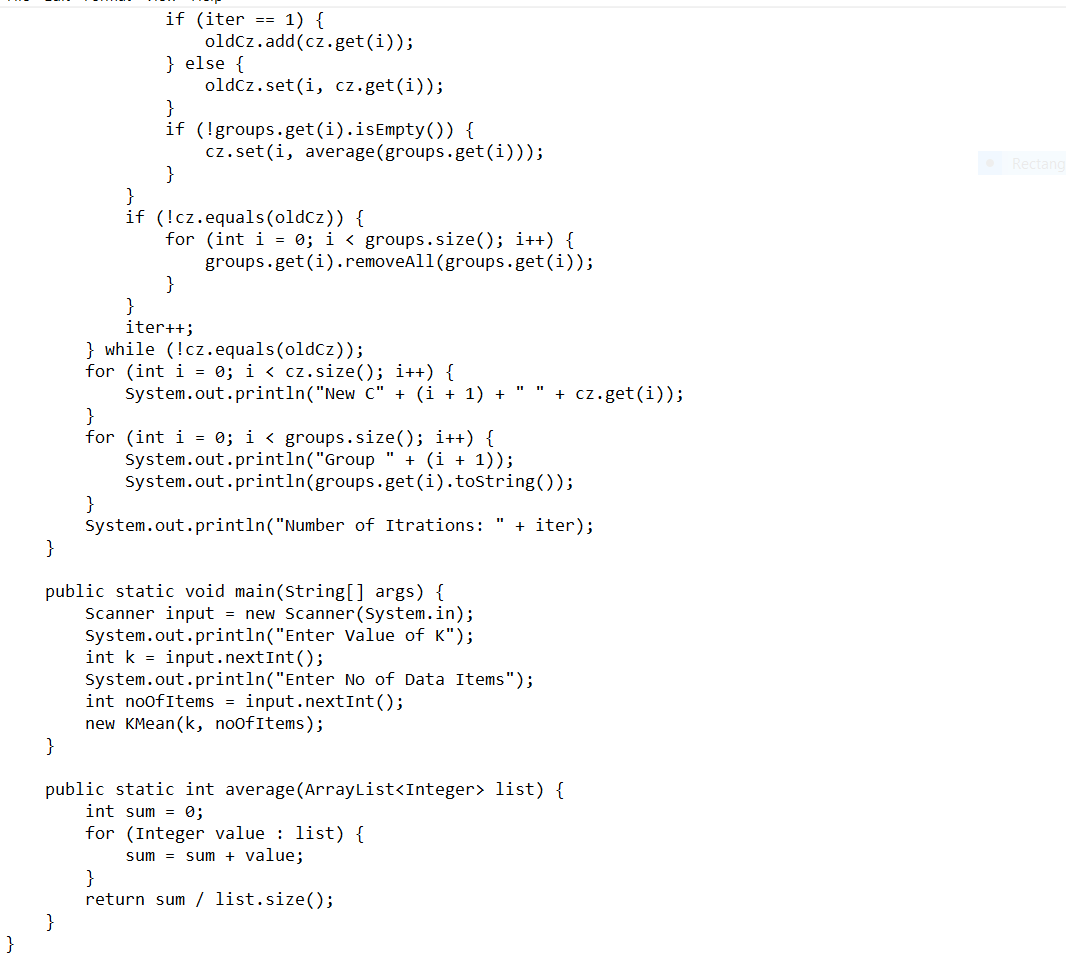
****

Fig 6.1 Data Set

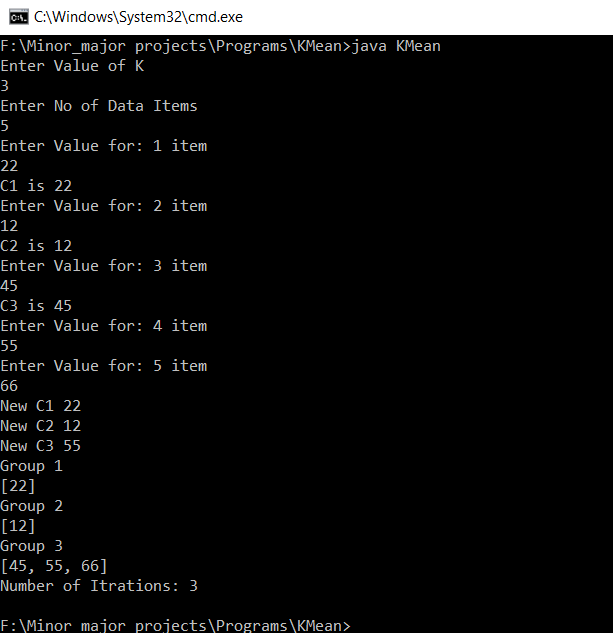
**6.2 K-MEANS IMPLEMENTATION**

1) CODE



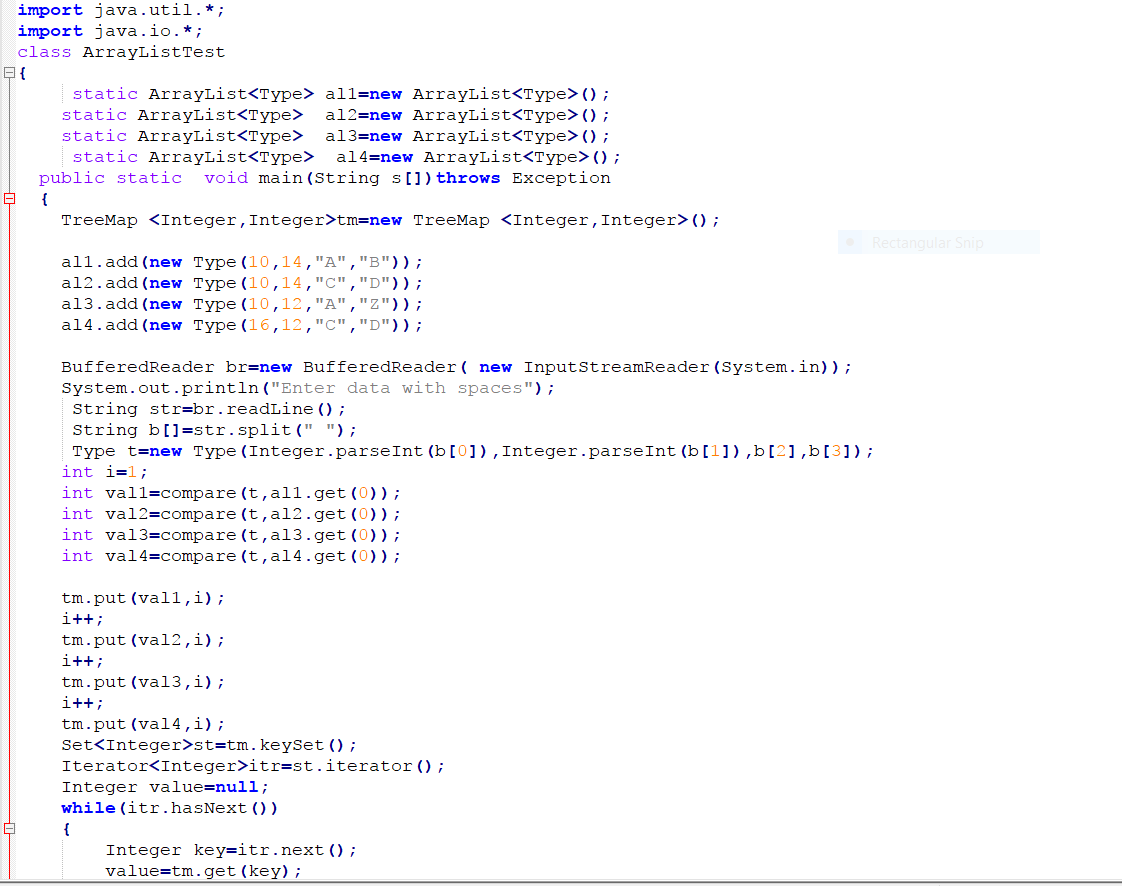


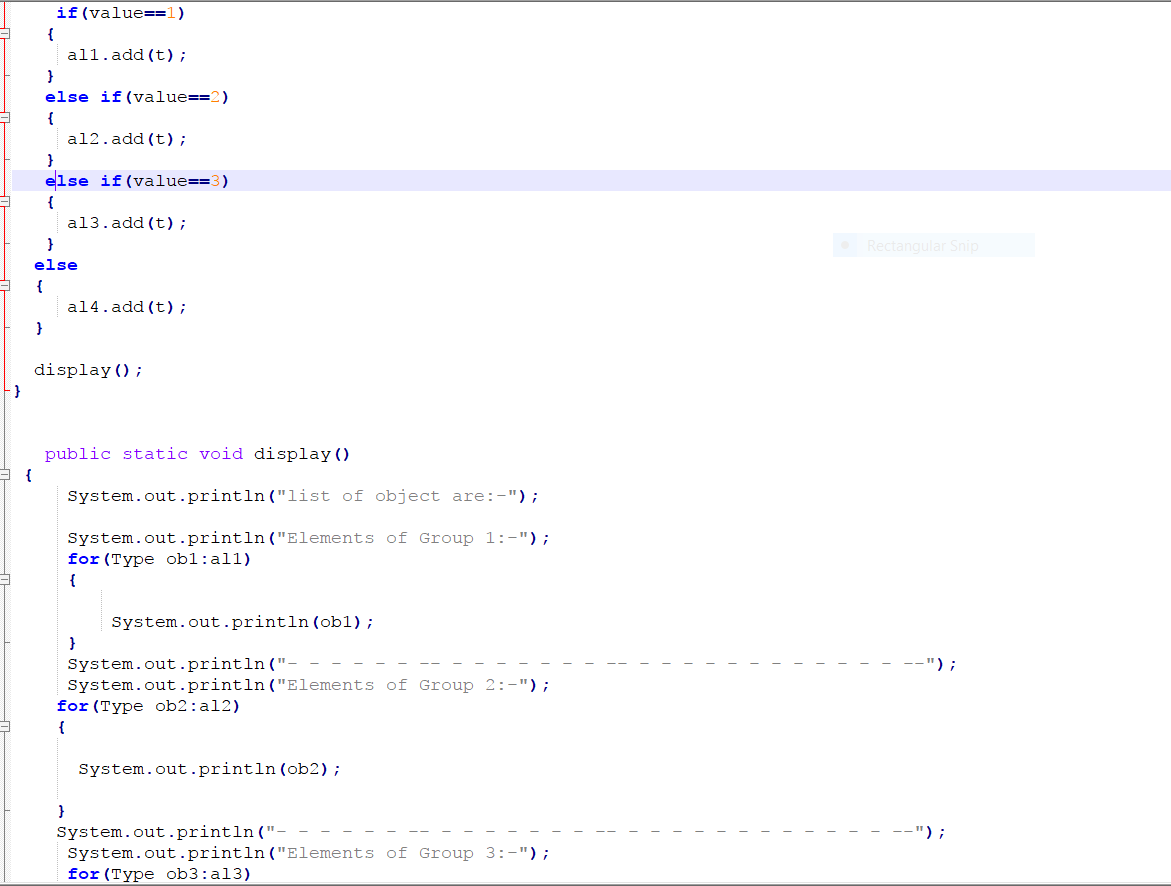
2) OUTPUT



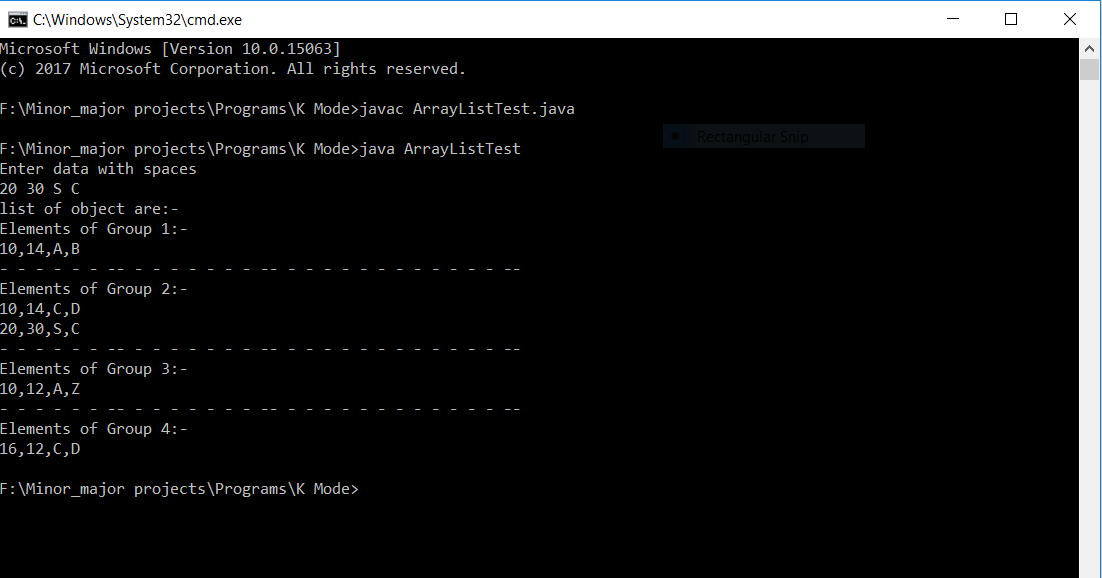
**6.3 K-MODE IMPLEMENTATION**

1) CODE

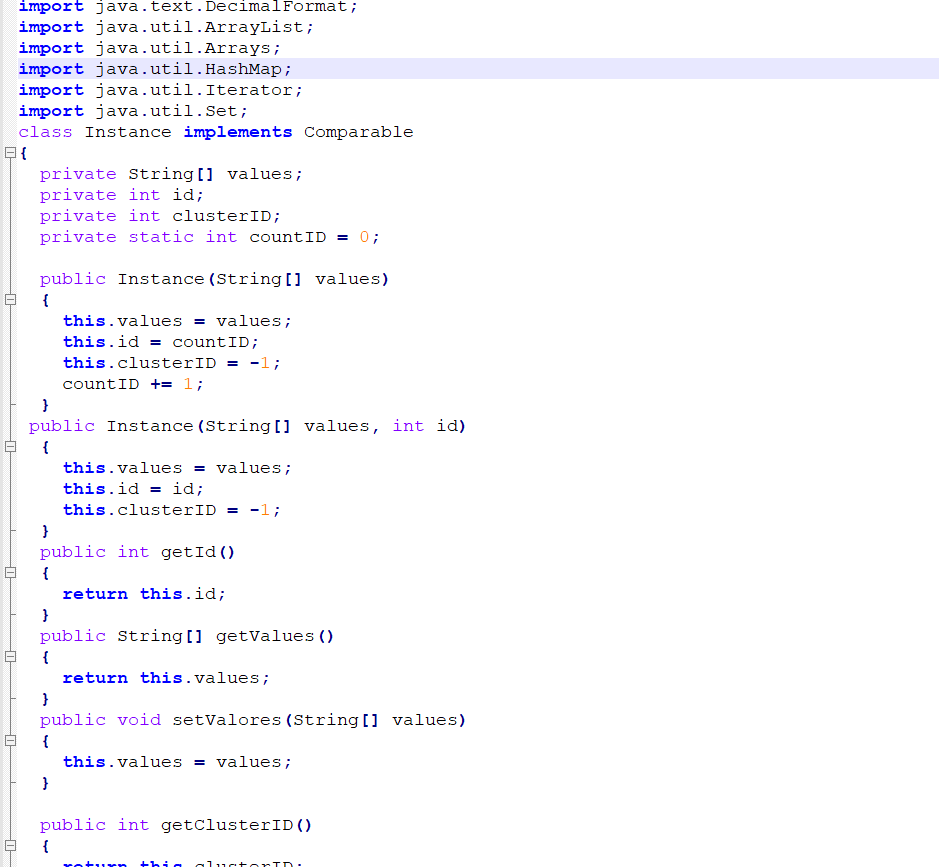
****

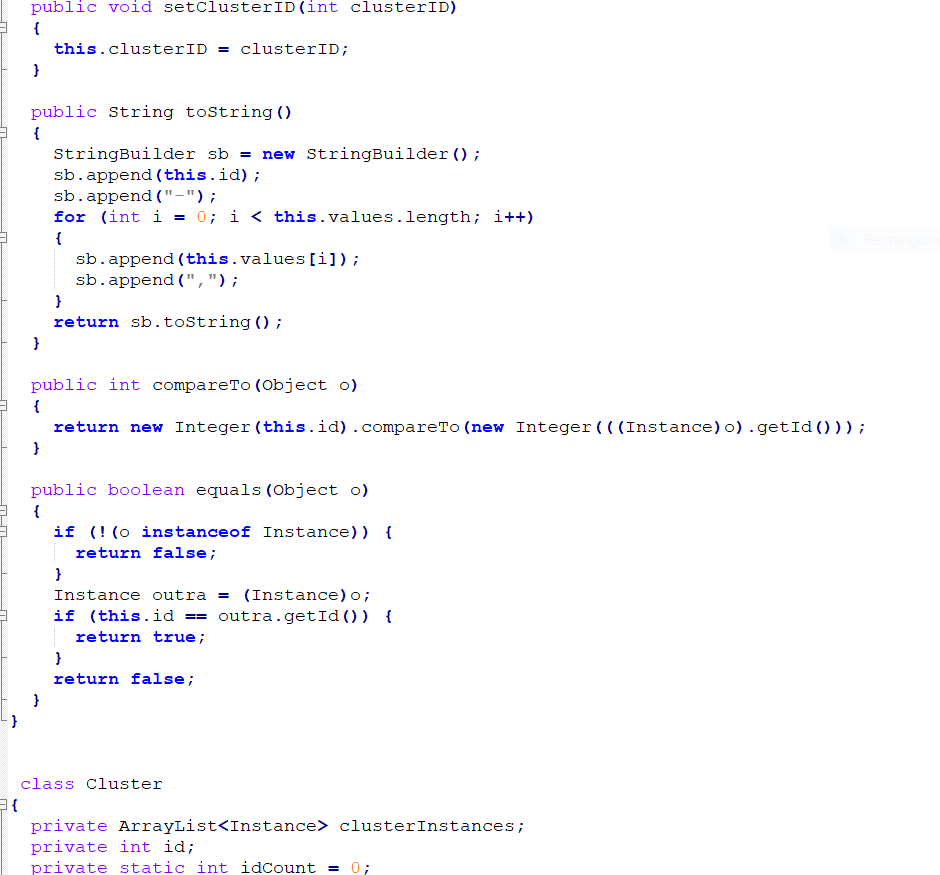
****

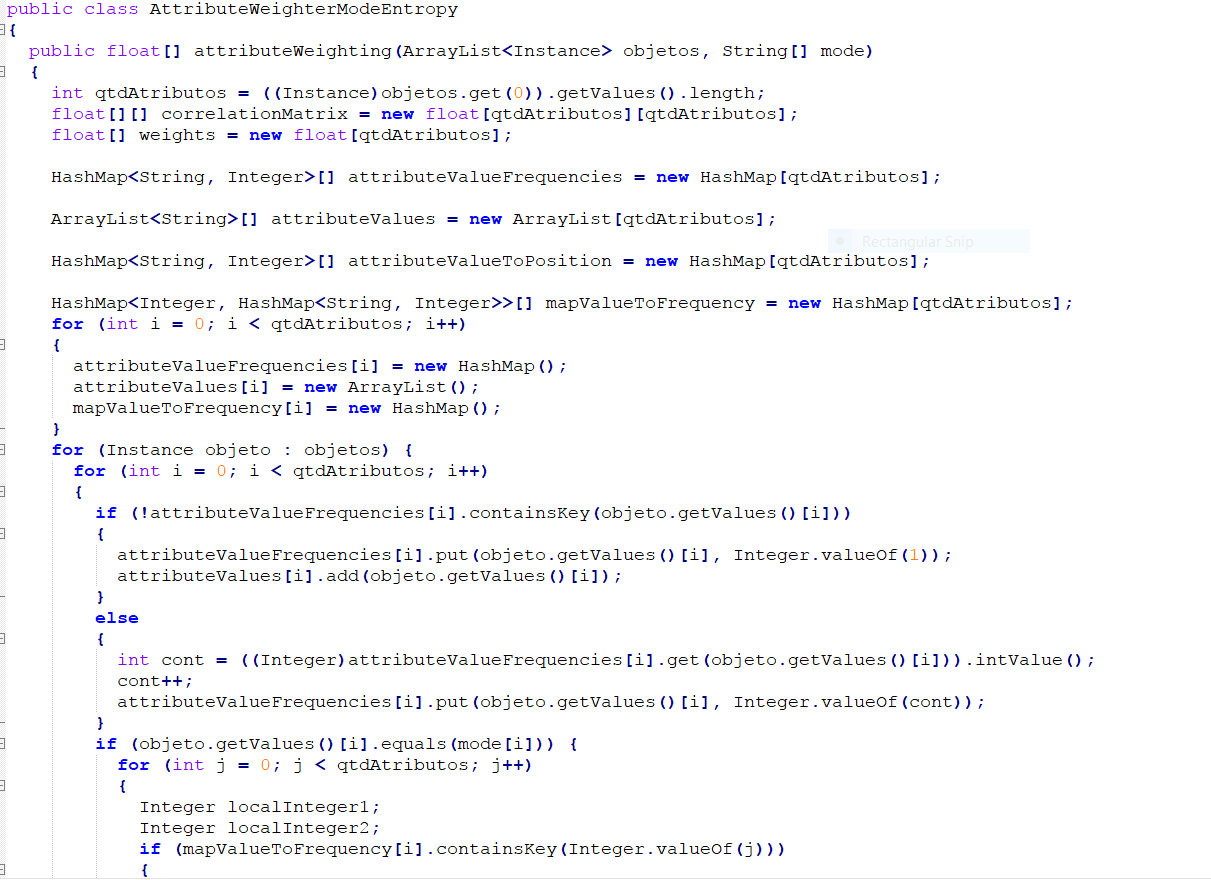
2) OUTPUT

****

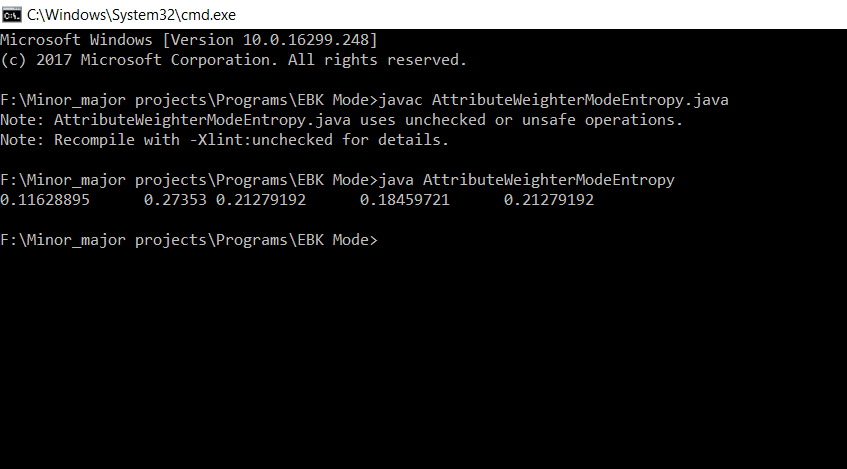
**6.4 EBK-MODE IMPLEMENTATION**

****

****

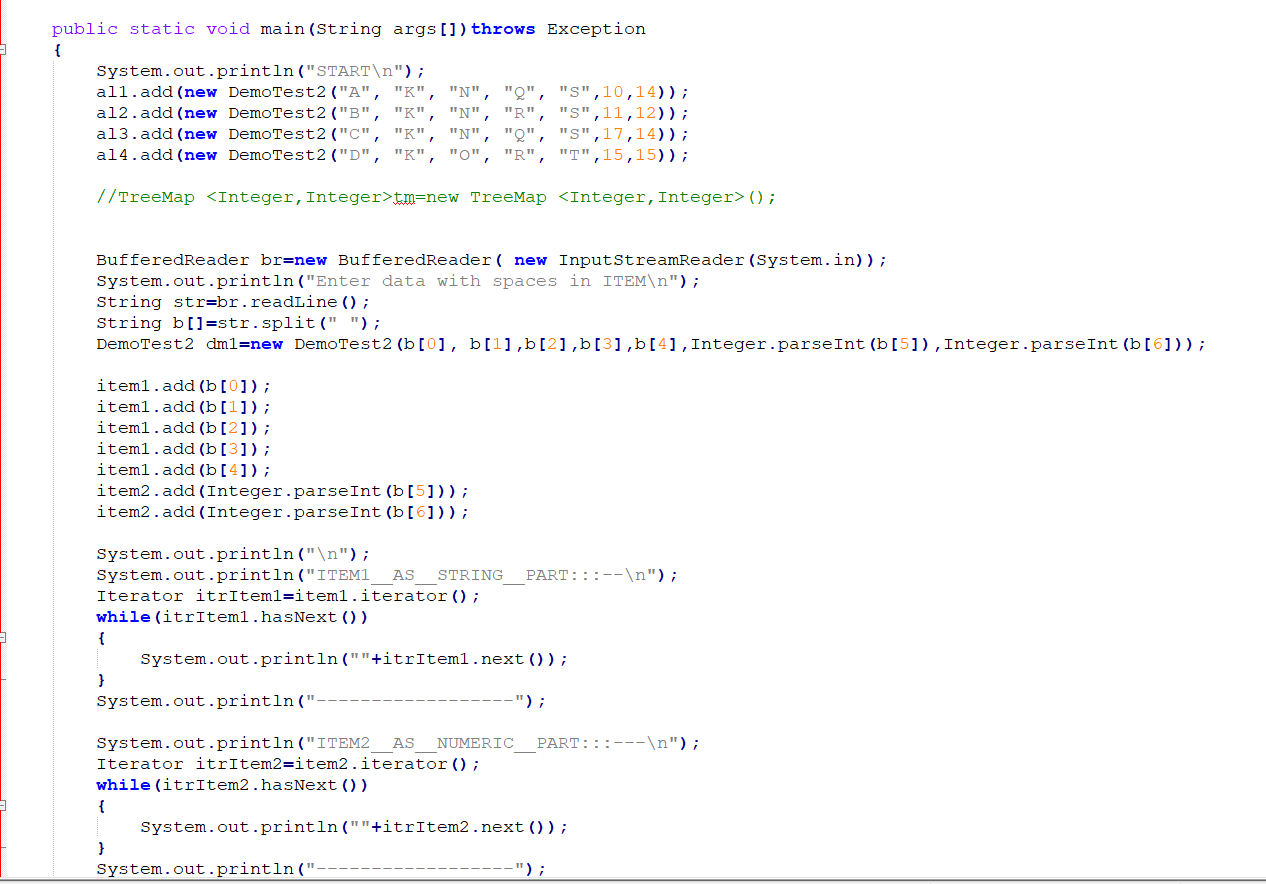
****

2) OUTPUT

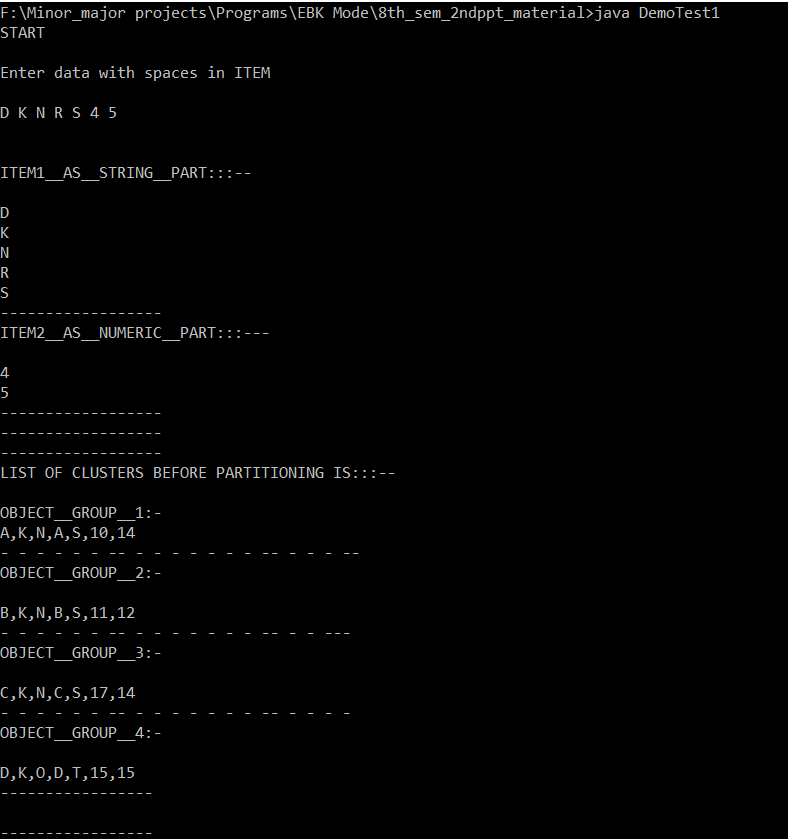


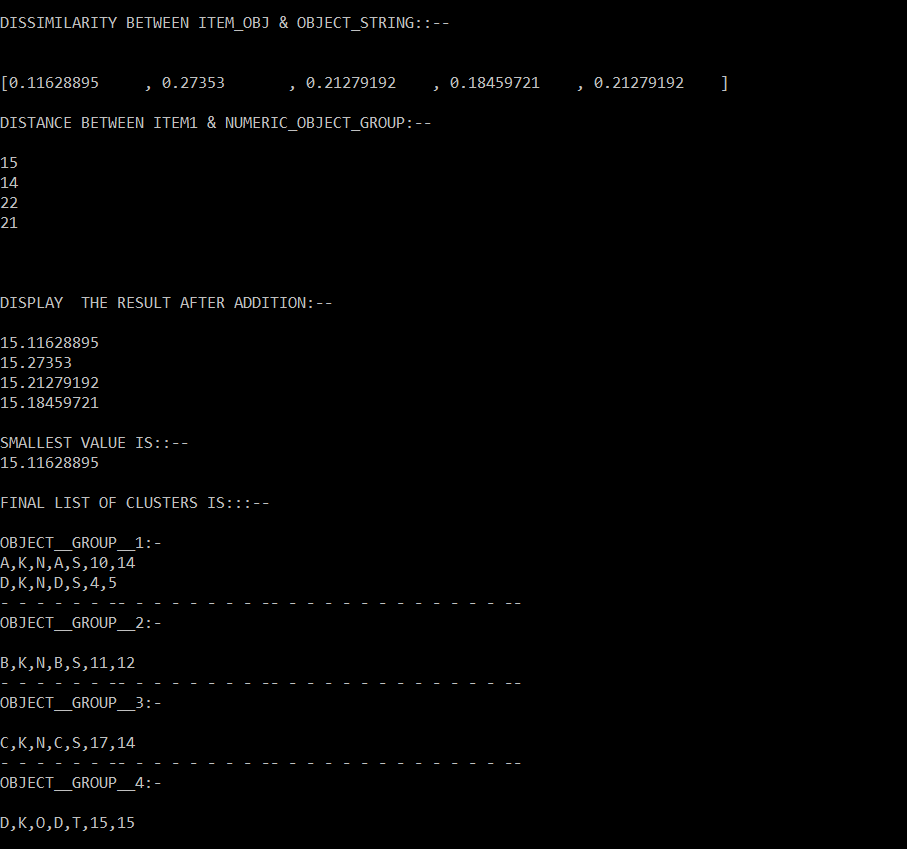
**6.5 EBK-MODE IMPLEMENTATION FOR MIXED DATA**

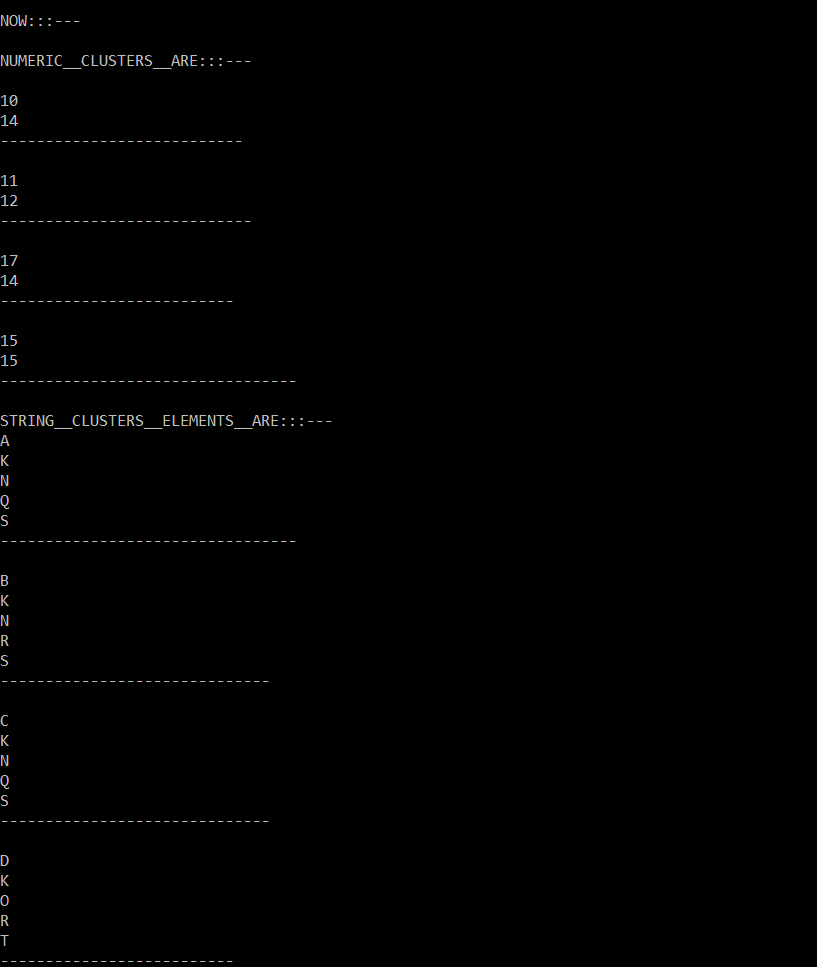
1) CODE

****

2) OUTPUT







**CHAPTER 7 CONCLUSIONS**

**7.1 CONCLUSIONS**

#### During the process we implemented a subspace clustering algorithm for categorical data and mixed data called EBK-modes.

#### It modifies the basic k-modes by considering the entropy-based relevance index (ERI) as a measure of the relevance of each attribute in each cluster.

#### The ERI of a given attribute is inversely proportional to the average of the entropy projected to this attribute for each attribute value of the mode of a cluster.

#### Our experiments have shown that the EBK-modes has a performance that is comparable to the performance of the state-of-the-art algorithms.

**REFERENCES**

**[1]** D. Barbar´a, Y. Li, and J. Couto, “Coolcat: an entropybased algorithm for categorical clustering,” in Proceedings of the eleventh international conference on Information and

knowledge management. ACM, 2002, pp. 582–589.

**[2]** P. Andritsos and P. Tsaparas, “Categorical data clustering,”in Encyclopedia of Machine Learning. Springer, 2010, pp.154–159.

**[3]** J. L. Carbonera and M. Abel, “Categorical data clustering:a correlation-based approach for unsupervised attribute weighting,” in Proceedings of ICTAI 2014, 2014.

**[4]** L. Bai, J. Liang, C. Dang, and F. Cao, “A novel attributeweighting algorithm for clustering high-dimensional categorical data,” Pattern Recognition, vol. 44, no. 12, pp. 2843–2861,2011.

**[5]** G. Gan and J. Wu, “Subspace clustering for high dimensional categorical data,” ACM SIGKDD Explorations Newsletter,vol. 6, no. 2, pp. 87–94, 2004.

**[6]** M. J. Zaki, M. Peters, I. Assent, and T. Seidl, “Clicks: An effective algorithm for mining subspace clusters in categorical datasets,” Data & Knowledge Engineering, vol. 60, no. 1, pp.

51–70, 2007.

**[7]** E. Cesario, G. Manco, and R. Ortale, “Top-down parameterfree Clustering of high-dimensional categorical data,” Knowledge and Data Engineering, IEEE Transactions on, vol. 19,

no. 12, pp. 1607–1624, 2007.

**[8]** H.-P. Kriegel, P. Kr¨oger, and A. Zimek, “Subspace clustering,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 2, no. 4, pp. 351–364, 2012.

**[9]** L. Jing, M. K. Ng, and J. Z. Huang, “An entropy weighting kmeans algorithm for subspace clustering of high-dimensional sparse data,” Knowledge and Data Engineering, IEEE Transactionson, vol. 19, no. 8, pp. 1026–1041, 2007.

**[10]** A. Keller and F. Klawonn, “Fuzzy clustering with weighting of data variables,” International Journal of Uncertainty,Fuzziness and Knowledge-Based Systems, vol. 8, no. 06, pp.

735–746, 2000.

**[11]** F. Cao, J. Liang, D. Li, and X. Zhao, “A weighting kmodes algorithm for subspace clustering of categorical data,”Neurocomputing, vol. 108, pp. 23–30, 2013.

**[12]** E. Y. Chan, W. K. Ching, M. K. Ng, and J. Z. Huang,“An optimization algorithm for clustering using weighted Dissimilarity measures,” Pattern recognition, vol. 37, no. 5,pp. 943–952, 2004.

**[13]** Z. He, X. Xu, and S. Deng, “Attribute value weighting in kmodes clustering,” Expert Systems with Applications, vol. 38,no. 12, pp. 15 365–15 369, 2011.

**[14]** C. E. Shannon, “A mathematical theory of communication,”ACM SIGMOBILE Mobile Computing and Communications Review, vol. 5, no. 1, pp. 3–55, 2001.

**[15]** Z. Huang, “Extensions to the k-means algorithm for clustering large data sets with categorical values,” Data mining and knowledge discovery, vol. 2, no. 3, pp. 283–304, 1998.

**[16]** B. Larsen and C. Aone, “Fast and effective text mining using linear-time document clustering,” in Proceedings of the fifth ACM SIGKDD international conference on Knowledge

discovery and data mining. ACM, 1999, pp. 16–22.

Data Set: Ucimachinelearning.Org