

# Cluster Center Initialization for Categorical Data Using Multiple Attribute Clustering

Shehroz S. Khan<sup>1</sup> Amir Ahmad<sup>2</sup>

<sup>1</sup>David R. Cheriton School of Computer Science  
University of Waterloo, Canada

<sup>2</sup>King Abdulaziz University  
Rabigh, Saudi Arabia

Introduction

K-Modes Clustering

Cluster Center Initialization

Proposed Approach

Results

Conclusions

# Clustering

- ▶ Unsupervised Learning
- ▶ Homogenous groups
- ▶ Diverse Application
  - ▶ Web Documentation
  - ▶ Image Analysis
  - ▶ Medical Analysis ...
- ▶ Types
  - ▶ Hierarchical -  $O(N^2)$ 
    - ▶ Agglomerative
    - ▶ Divisive
  - ▶ Partitional -  $O(N)$
  - ▶ Density / Distribution based ...

# Formulation

- ▶ K-means
  - ▶ Process large numeric datasets
  - ▶ Simple and Efficient
  - ▶ Fails to handle datasets with categorical attributes because it minimizes the cost function by calculating *means*

# Formulation

- ▶ K-means
  - ▶ Process large numeric datasets
  - ▶ Simple and Efficient
  - ▶ Fails to handle datasets with categorical attributes because it minimizes the cost function by calculating *means*
- ▶ K-modes [Huang, 1997]
  - ▶ new dissimilarity measure

$$d(X, Y) = \sum_{j=1}^m \delta(x_j, y_j) \quad (1)$$

$$\text{where } \delta(x_j, y_j) = \begin{cases} 0 & (x_j = y_j) \\ 1 & (x_j \neq y_j) \end{cases}$$

- ▶ replaces *means* of clusters with *modes*,
- ▶ use a frequency based method to update *modes* in the clustering process to minimize the cost function

# Algorithm

1. Create  $K$  clusters by randomly choosing data objects and select  $K$  initial cluster centers, one for each of the cluster.
2. Allocate data objects to the cluster whose cluster center is nearest to it according to the objective function.
3. Update the  $K$  clusters based on allocation of data objects and compute  $K$  new modes of all clusters.
4. Repeat step 2 to 3 until no data object has changed cluster membership or any other predefined criterion is fulfilled.

## Advantages and Limitations

- ▶ Achieves convergence with linear time complexity
- ▶ Faster than the K-means algorithm [Huang, 1998]
- ▶ Assumes that the number of clusters,  $K$ , is known in advance
- ▶ Falls into problems when clusters are of differing sizes, density and non-globular shapes

## Advantages and Limitations

- ▶ Achieves convergence with linear time complexity
- ▶ Faster than the K-means algorithm [Huang, 1998]
- ▶ Assumes that the number of clusters,  $K$ , is known in advance
- ▶ Falls into problems when clusters are of differing sizes, density and non-globular shapes
- ▶ **Very sensitive to the choice of initial centers**



# Initialization Methods

- ▶ Random Initialization
  - ▶ Widely used, Simple but non-repeatable results
  - ▶ Does not guarantee unique clustering
  - ▶ Improper choice may yield highly undesirable cluster structures

# Initialization Methods

- ▶ Random Initialization
  - ▶ Widely used, Simple but non-repeatable results
  - ▶ Does not guarantee unique clustering
  - ▶ Improper choice may yield highly undesirable cluster structures
- ▶ Other Methods of Initialization
  - ▶ Non-linear in time complexity with respect to the number of data objects
  - ▶ Initial modes are not fixed and possess some kind of randomness in the computation steps
  - ▶ Dependent on the presentation of order of data objects

# Multiple Attribute Clustering Approach

Based on the following experimental observations

1. Some of the data objects are very similar to each other and they have same cluster membership irrespective of the choice of initial cluster centers [Khan and Ahmad, 2004].
2. There may be some attributes in the dataset whose number of attribute values are less than or equal to  $K$ . Due to fewer attribute values per cluster, these attributes shall have higher discriminatory power and will play a significant role in deciding the initial modes as well as the cluster structures. We call them as *Prominent Attributes (P)* .

# Main Idea

- ▶ For every prominent attribute, partition the data based on its attribute values  $j$

# Main Idea

- ▶ For every prominent attribute, partition the data based on its attribute values  $j$
- ▶ Divide the dataset into  $j$  clusters on the basis of these  $j$  attribute values such that data objects of  $i^{th}$  attribute with different values fall into different clusters.

# Main Idea

- ▶ For every prominent attribute, partition the data based on its attribute values  $j$
- ▶ Divide the dataset into  $j$  clusters on the basis of these  $j$  attribute values such that data objects of  $i^{th}$  attribute with different values fall into different clusters.
- ▶ Compute the modes, use them as initial modes, cluster data and generate a *cluster string* that contains the respective cluster allotment labels of the full data.

# Main Idea

- ▶ For every prominent attribute, partition the data based on its attribute values  $j$
- ▶ Divide the dataset into  $j$  clusters on the basis of these  $j$  attribute values such that data objects of  $i^{th}$  attribute with different values fall into different clusters.
- ▶ Compute the modes, use them as initial modes, cluster data and generate a *cluster string* that contains the respective cluster allotment labels of the full data.
- ▶ A number of cluster strings are generated that represent different partition views of the data. If needed, merge the distinct similar cluster strings into  $K$  partitions

# Main Idea

- ▶ For every prominent attribute, partition the data based on its attribute values  $j$
- ▶ Divide the dataset into  $j$  clusters on the basis of these  $j$  attribute values such that data objects of  $i^{th}$  attribute with different values fall into different clusters.
- ▶ Compute the modes, use them as initial modes, cluster data and generate a *cluster string* that contains the respective cluster allotment labels of the full data.
- ▶ A number of cluster strings are generated that represent different partition views of the data. If needed, merge the distinct similar cluster strings into  $K$  partitions
- ▶ *Cluster strings* within each  $K$  clusters are replaced by the corresponding data objects and modes of every  $K$  cluster is computed that serves as the initial centers for the K-modes



# Conditions

- ▶ Prominent Attributes
  - ▶ If  $\#P > 0$ , then use only *Prominent* attributes
  - ▶ If  $\#P = 0$ , then use all attributes

# Conditions

- ▶ Prominent Attributes
  - ▶ If  $\#P > 0$ , then use only *Prominent* attributes
  - ▶ If  $\#P = 0$ , then use all attributes
- ▶ Distinct Cluster Strings,  $K'$  (distinguishable clusters)
  1.  $K' > K \rightarrow$  Merge similar distinct cluster string and compute initial modes
  2.  $K' = K \rightarrow$  Distinct cluster strings matches the desired number of clusters in the data.
  3.  $K' < K \rightarrow$ 
    - ▶ when the partitions created based on the attribute values of attributes group the data in the same clusters every time
    - ▶ when the attribute values of all attributes follow almost same distribution
    - ▶ probably the chosen  $K$  does not resemble with the natural grouping and it should be changed

# Scenarios

- ▶ Sort and choose top  $K$

# Scenarios

- ▶ Sort and choose top  $K$
- ▶ Hierarchical clustering
  - ▶  $K'$  cluster strings are less than  $N$
  - ▶ Choose the most frequent  $N^{0.5}$  distinct cluster strings
  - ▶ Log Linear Complexity
  - ▶ Infrequent cluster strings can be considered as outliers or boundary cases

# Scenarios

- ▶ Sort and choose top  $K$
- ▶ Hierarchical clustering
  - ▶  $K'$  cluster strings are less than  $N$
  - ▶ Choose the most frequent  $N^{0.5}$  distinct cluster strings
  - ▶ Log Linear Complexity
  - ▶ Infrequent cluster strings can be considered as outliers or boundary cases
- ▶ Choice of Attributes
  - ▶ For  $\#P=0 \rightarrow$  increased number of distinct cluster strings
  - ▶ Choosing  $\sqrt{N}$  cluster strings may result in loss of information

# Scenarios

- ▶ Sort and choose top  $K$
- ▶ Hierarchical clustering
  - ▶  $K'$  cluster strings are less than  $N$
  - ▶ Choose the most frequent  $N^{0.5}$  distinct cluster strings
  - ▶ Log Linear Complexity
  - ▶ Infrequent cluster strings can be considered as outliers or boundary cases
- ▶ Choice of Attributes
  - ▶ For  $\#P=0 \rightarrow$  increased number of distinct cluster strings
  - ▶ Choosing  $\sqrt{N}$  cluster strings may result in loss of information
- ▶ Time Complexity
  - ▶ Log Linear  $\rightarrow$  worst case

## Effect of Choosing Different Number of Attributes

Dataset	Proposed		Vanilla		$\sqrt{N}$
	#P	#CS	#A	#CS	
Soybean	20	21	35	25	7
Zoo	16	7	17	100	11
Breast-Cancer	9	355	9	355	27
Lung-Cancer	54	32	56	32	6
Mushroom	5	16	22	683	91

## Performance

Table: Breast Cancer data

	Random	Wu	Cao	Proposed
<b>AC</b>	0.8364	0.9113	0.9113	<b>0.9127</b>
<b>PR</b>	0.8699	<b>0.9292</b>	<b>0.9292</b>	<b>0.9292</b>
<b>RE</b>	0.7743	0.8773	0.8773	<b>0.8783</b>

Table: Zoo data

	Random	Wu	Cao	Proposed
<b>AC</b>	0.8356	0.8812	0.8812	<b>0.891</b>
<b>PR</b>	0.8072	<b>0.8702</b>	<b>0.8702</b>	0.7302
<b>RE</b>	0.6012	0.6714	0.6714	<b>0.8001</b>



## Performance

Table: Mushroom data

	Random	Wu	Cao	Proposed
<b>AC</b>	0.7231	0.8754	0.8754	<b>0.8815</b>
<b>PR</b>	0.7614	<b>0.9019</b>	<b>0.9019</b>	0.8975
<b>RE</b>	0.7174	0.8709	0.8709	<b>0.8780</b>

Table: Lung Cancer data

	Random	Wu	Cao	Proposed
<b>AC</b>	<b>0.5210</b>	0.5	0.5	0.5
<b>PR</b>	0.5766	0.5584	0.5584	<b>0.6444</b>
<b>RE</b>	0.5123	0.5014	0.5014	<b>0.5168</b>

# Comparison





- ▶ Other Approaches
  - ▶ Random Initialization → non-repeatable and poor results
  - ▶ Wu et al [Wu et al., 2007] → induces random selection of data points
  - ▶ Cao et al. [Cao et al., 2009] → quadratic complexity

# Comparison

- ▶ Other Approaches
  - ▶ Random Initialization → non-repeatable and poor results
  - ▶ Wu et al [Wu et al., 2007] → induces random selection of data points
  - ▶ Cao et al. [Cao et al., 2009] → quadratic complexity
- ▶ Proposed Approach
  - ▶ Fixed Initial Clusters, Repeatable results
  - ▶ Independent of order of data presentation
  - ▶ Better Performance
  - ▶ Worst Case Complexity – Log Linear

- ▶ Results attained by the K-modes algorithm depends intrinsically on the choice of random initial cluster centers
- ▶ Proposed a Multiple attribute clustering approach for finding fixed initial modes
- ▶ Extension – Finding out the natural number of clusters present in the data?

THANKS

-  Cao, F., Liang, J., and Bai, L. (2009).  
A new initialization method for categorical data clustering.  
*Expert Systems and Applications*, 36:10223–10228.
-  Huang, Z. (1997).  
A fast clustering algorithm to cluster very large categorical data sets in data mining.  
*In Research Issues on Data Mining and Knowledge Discovery*.
-  Huang, Z. (1998).  
Extensions to the k-means algorithm for clustering large data sets with categorical values.  
*Data Min. Knowl. Discov.*, 2(3):283–304.
-  Khan, S. S. and Ahmad, A. (2004).  
Cluster center initialization algorithm for k-means clustering.  
*Pattern Recognition Letters*, 25:1293–1302.



Wu, S., Jiang, Q., and Huang, J. Z. (2007).

A new initialization method for clustering categorical data.

In *Proceedings of the 11th Pacific-Asia conference on Advances in knowledge discovery and data mining, PAKDD'07*, pages 972–980, Berlin, Heidelberg. Springer-Verlag.