

# An Accurate Grid -based PAM Clustering Method for Large Dataset

Faisal Bin Al Abid  
Department of Computer  
Science and Engineering  
Islamic University of  
Technology  
Bangladesh.

M.A. Mottalib  
Department of Computer  
Science and Engineering  
Islamic University of  
Technology  
Bangladesh.

## ABSTRACT

Clustering is the procedure to group similar objects together. Several algorithms have been proposed for clustering. Among them, the K-means clustering method has less time complexity. But it is sensitive to extreme values and would cause less accurate clustering of the dataset. However, K-medoids method does not have such limitations. But this method uses user-defined value for K. Therefore, if the number of clusters is not chosen correctly, it will not provide the natural number of clusters and hence the accuracy will be minimized. In this paper, we propose a grid based clustering method that has higher accuracy than the existing K-medoids algorithm. Our proposed Grid Multi-dimensional K-medoids (GMK) algorithm uses the concept of cluster validity index and it is shown from the experimental results that the new proposed method has higher accuracy than the existing K-medoids method. The object space is quantized into a number of cells, and the distance between the intra cluster objects decrease which contributes to the higher accuracy of the proposed method. Therefore, the proposed approach has higher accuracy and provides natural clustering method which scales well for large dataset.

## General Terms

Data Mining.

## Keywords

Medoid, Grid, ADULT Dataset, Partitioning, Cluster validity index, Dense grid, Outlier detection, accuracy.

## 1. INTRODUCTION

Data Mining is the procedure of non-trivial extraction of implicit, previously unknown, and potentially helpful information from data [1]. Commonly, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks illustrate the general properties of the data in the database i.e. descriptive task finds the human-interpretable patterns that describe the data. Predictive mining tasks perform inference on the existing data in order to make predictions. Clustering is one of the major descriptive data mining tasks. As mentioned, clustering is partitioning of data into groups of analogous objects. Representing the dataset by fewer clusters loses certain fine details, but achieves simplification [1]. Data modeling puts clustering in a historical viewpoint rooted in mathematics, statistics, and numerical analysis. Clustering can be viewed as a density evaluation problem. From a machine learning viewpoint clusters correspond to hidden patterns, the exploration for

clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays a marvelous role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.

Clustering is the topic of active research in several fields such as statistics, pattern recognition, and machine learning. Data mining adds to clustering the problems of very large datasets with very many attributes of different types. This enforces sole computational

prerequisites on relevant clustering algorithms. A variety of algorithms have recently emerged that meet these requirements and were successfully applied to real-life data mining problems. Clustering in data mining was brought to life by intense developments in information retrieval and text mining, spatial database applications, for example, GIS or astronomical data [2], sequence and heterogeneous data analysis [3], Web applications [4], DNA analysis in computational biology [5], and many others.

K-means clustering method is a popular clustering algorithm since it has less time complexity. However, it suffers from sensitivity of outliers which may distort the distribution of data due to the extreme values. Due to the sensitivity of outlier in K-means, we are dealing with K-medoids clustering method. Representation by K-medoids has two advantages:

- (i) It presents no constraints on attributes types.
- (ii) The preference of medoids is dictated by the location of a predominant i.e. major fraction of points inside a cluster and, therefore, it is lesser sensitive to the presence of outliers and noise.

On the other hand, the disadvantages of K-medoids are:

- (i) It has the value of K used as user defined.
- (ii) It does not scale well for large data set.

In this paper, we mainly concentrated on eliminating these disadvantages using grid clustering GMK approach. In this approach, we first partition the grid and put the data values inside the grid. Each grid is considered as a cluster and bottom-up approach is used to find the center of a cluster. Thus, the user will not have to specify the number of clusters

and it does not need an iterative approach to deal with large dataset. Thus it will provide the natural clusters with less time complexity. The rest of the paper is organized as follows: Section 2 presents the taxonomy of different clustering methods. Section 3 and 4 describes the variants of K-medoids method and the K-medoids method respectively. Section 5 presents the proposed Grid Multidimensional K-medoids (GMK) method. Section 6 illustrates the experimental results using ADULT dataset. Finally, section 7 contains the concluding remarks.

## 2. TAXONOMY OF CLUSTERING

There are several well-known clustering algorithms; different clustering algorithms may provide different clusters. The most well known clustering algorithms are hierarchical clustering, density based clustering, grid based clustering, model based clustering and partition based clustering. Hierarchical clustering constructs a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram. Every cluster node contains child clusters; sibling clusters partition the points covered by their common parent. Such an approach allows discovering data on different levels of granularity. Hierarchical clustering methods are categorized into agglomerative (bottom-up) and divisive (top-down) methods [6]. An agglomerative clustering starts with one-point (singleton) clusters and recursively merges two or more most appropriate clusters. A divisive clustering starts with one cluster of all data points and recursively splits the most appropriate cluster.

An open set in the Euclidean space can be divided into a set of its connected components. The implementation of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary. They are closely related to a point's nearest neighbors. A cluster, defined as a connected dense component, grows in any direction that density leads. Therefore, density-based algorithms are able to discover clusters of arbitrary shapes. Also this provides a natural protection against outliers or noise. Since density-based algorithms require a metric space, the natural setting for them is spatial data clustering [7]. To make computations feasible, some index of data is constructed (such as R\*-tree). This is a topic of active research. Classic indices were helpful only with reasonably low-dimensional data. The algorithm DENCLUE that, in fact, is a mixture of a density-based clustering and a grid-based preprocessing is lesser affected by data dimensionality.

Grid based clustering methods are used for multi resolution data structure. It is used to quantize the object space into a finite number of cells that form a grid structure on which all the actions are to be performed. The main advantage of grid clustering is faster processing time, which is typically free from the number of data objects, yet dependent on the number of cells in each dimension in the quantized space. Some typical example of grid based clustering are STING (Statistical information grid) which represents and explores grid information stored in grid cells and processes fast than other conventional clustering [8]. Data partitioning clustering algorithms divide data into several subsets. Because checking

all probable subset systems is computationally impossible, certain greedy heuristics are used in the form of iterative optimization. Specifically, this means different relocation schemes that iteratively reassign points between the  $k$  clusters. Unlike traditional hierarchical methods, in which clusters are not revisited after being constructed, relocation algorithms gradually improve clusters. With appropriate dataset, this results in high quality clusters. One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found. Another approach starts with the definition of objective function depending on a partition, computation of objective function becomes linear in  $N$  (and in a number of clusters  $K \ll N$ ). Depending on how representatives are constructed, iterative optimization partitioning algorithms are subdivided into K-medoids and K-means methods. K-medoids is the most appropriate data point within a cluster that represents it.

## 3. VARIANTS OF K-MEDOIDS

The One of the most well-known versions of K-medoids are PAM (Partitioning Around Medoids). PAM is iterative optimization that combines relocation of points between perspective clusters with re-nominating the points as potential medoids. The guiding principle for the process is the effect on an objective function, which, obviously, is a costly strategy. CLARA uses several samples, each with  $40+2K$  points, which are each subjected to PAM. The whole dataset is assigned to resulting medoids, the objective function is computed, and the best system of medoids is retained. CLARA is used to deal with very large data set. Further progress is associated with Ng & Han who introduced the algorithm CLARANS (Clustering Large Applications based upon Randomized Search) in the context of clustering in spatial databases [9]. It uses sample with some randomness at each step of the search. Theoretically the clustering process can be viewed as a search through a graph, where each node is a potential solution (a set of  $k$  medoids). Two nodes are neighbors (that is, connected by arc in the graph) if their sets differ by only one object. Each node can be assigned a cost. PAM searches and examines all of the neighbors of the current node in its search for a minimum cost. CLARA has time complexity  $O(Ks^2 + K(n-K))$ , CLARANS has time complexity  $O(N^2)$ . As mentioned above, we will focus our view on the basic K-medoids method, because if this proposed method works well, it will work well for CLARA and CLARANS that deals with larger data set. An improved K-medoids method has been proposed based on cluster validity index  $V_{xb}$  as mentioned in subsection of 5 below. This improved version of K-medoids method chooses the optimum cluster for clustering but the time complexity of the improved K-medoids method is too high. The Xie-Beni index used to determine the cluster validity index is the ratio of the average intra-cluster compactness to inter-cluster separation between the clusters. In this paper, we will use the Xie-Beni index in order to compare the accuracy of the existing K-medoids method with the proposed GMK method.

#### 4. THE K-MEDOIDS METHOD

The most common realization of *K-medoids* clustering is the Partitioning Around Medoids (PAM) algorithm and is as follows:

1. **Initialize:** randomly choose *K* of the *n* data points as the medoids
2. Associate each data point to the *closest medoid*.  
*("closest" here is defined using any valid distance metric, most commonly Euclidean distance, Manhattan distance or Minkowski distance)*
3. For each medoid *m*
4. For each non-medoid data point
  - (i) Swap *m* and *o*
  - (ii) compute the total cost of the configuration
5. Select the configuration with the lowest cost.
6. **Repeat** steps 2 to 4 until there is no change in the medoid.

The flowchart in figure 1 describes about the conventional *K-medoids* method:

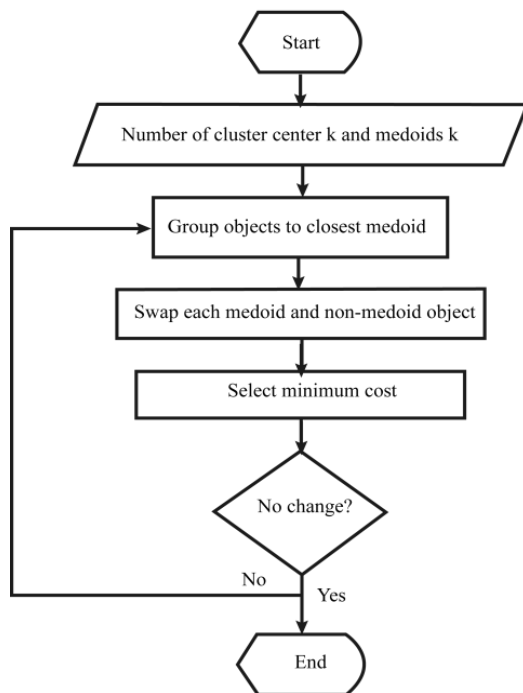


Figure 1: Flowchart of *K-medoids* method

#### 5. PROPOSED GMK CLUSTERING METHOD

The concept is taken from GK-means method to partition the grid [10].The algorithmic procedure of GMK is subdivided

into five main parts: Getting special grids, Detection of outlier, cell merging, center selection of merged clusters, determining the accuracy of clustering. We divide the algorithm into sub-algorithms and describe each and every part of the sub-algorithms.

##### 5.1 Getting Special Grid

###### ALGORITHM1:

```

Get Grid (Point [] data set)
N := dataset.length
sigmaX := sqrt(N/m)
sigmaY := sqrt(N/m)
maxX := MIN_VALUE
maxY := MIN_VALUE
minX := MAX_VALUE
minY := MAX_VALUE
for (Point point : this.dataset)
    if(point.getX()[0] > maxX) then
        maxX := point.getX()[0]
    end if
    if(point.getX()[1] > this.maxY) then
        maxY := point.getX()[1]
    end if
    if(point.getX()[0] < minX) then
        minX := point.getX()[0]
    end if
    if(point.getX()[1] < this.minY) then
        minY := point.getX()[1]
    end if
    Lx := (maxX - minX)/sigmaX
    Ly := (maxY - minY)/sigmaY
    grid := [N][2]
end for
    
```

###### Description of algorithm1:

Get special grids based on the formula  $L_x = (\max_x - \min_x) / \sigma_x$  and  $\sigma_x = \sqrt{\frac{N}{m}}$  where  $L_x$  is the interval length in *x* dimension,  $\max_x$  is the maximum data value in *x* dimension,  $\min_x$  is the minimum data value in *x* dimension,  $\sigma_x$  is the number of segments in *x* dimension, *N*=total number of data points and *m* is the average number of data points in each grid. The interval for *y* dimension is calculated using the same formula. The objects are placed into the grid structure and the outlier of the objects are detected in the next step.

##### 5.2 Detection of outlier

###### ALGORITHM2:

```

Is Out( int w, int max) // Is used to detect the outlier
If ( 0.05>w/max)
    Outlier- Grid_cluster := cluster
else
    Cluster := cluster
end if
    
```

###### Description of algorithm2:

There is no hard and fast rule for the determination of outliers. In order to determine the outliers, first the cluster size of each

and every cell is computed. As stated above, Cluster size of a cell is Cluster (size) = Number of data points in a cell / maximum number of data points in all cells. If Cluster (size) of a cell is less than or equal to 5% of maximum number of data points in a cell the cell is not considered as outlier grid and the cell is not used for merging in order to produce the final cluster. But the outlier grid is kept in order to determine any kind of anomaly for future.

### 5.3 Cell merging

#### ALGORITHM3:

```

Merge () // For merging use the flood fill algorithm
Cell Cluster []
Insert Parent node into Q
While (!Q. empty)
    Cell Parent := Q. front ()
    Q. pop ()
    For all neighboring child of Parent
        Q. add (child)
    set child visited
end while

```

#### Description of algorithm3:

The neighboring cells which are not outliers are merged by searching the adjacent four neighboring cells or grids : Top, Bottom, Right and Left. The cells that are already merged are not considered for merging in future. The neighboring cells are merged using the flood fill algorithm[11] .

### 5.4 Center selection of merged cluster

```

Udist (Point a, Point b)
return
Math.sqrt (( a.getX()[0] -b.getX()[0])* (a.getX()[0] -
b.getX()[0]) + (a.getX()[1] - b.getX()[1])* (a.getX()[1] -
b.getX()[1]))
Repeat for j=0 to K by 1
    if (minDis > Udist(adult[i], adult[medoids[j]]))
        medIndex:= medoids[j]
        minDis:= Udist(adult[i], adult[medoids[j]])
    tmpSum := tmpSum +Udist(adult[i], adult[medIndex])
    tmpMedOfPoint[i] := medIndex;
end if
end for

```

#### Description of algorithm4:

For each and every merged neighboring cell or grid a cluster is formed. Inside each cluster, each point is used to calculate the distance between it to the rest of the points. The point which has the least cost using Euclidean distance is used as the centre of the cluster. The conventional K-medoids partitioning around method uses top-down approach where

the center is first taken randomly to calculate the cost. On the contrary, the bottom –up approach is used for the proposed method where after forming the merged cells, that is the cluster, the center is selected for the cluster.

### 5.5 Determining the accuracy of clustering

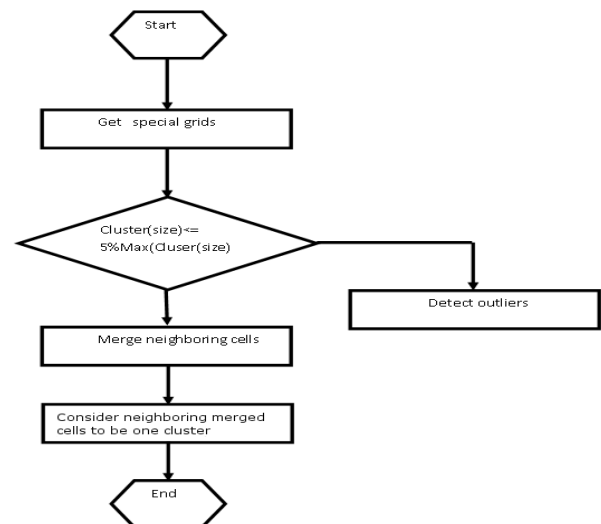
```

Repeat for i = 0 to result.size() by 1
    compactness := compactness+ Udist(result.get(i),
    centers.get(j-1))
Repeat for i = 0 to centers.size() by 1
    Repeat for j = 0 to centers.size() by 1
        if(i!=j)
            if(separation > Udist(centers.get(i),
            centers.get(j))
                separation:= Udist(centers.get(i), centers.get(j))
            end if
        end if
    end for
end for
Vxb := (compactness/N) / separation.

```

#### Description of algorithm5:

The definitions of compactness and separation is:  
 Compactness:  $1/n \sum \sum \|a_k - p_l\|^2$  (1)  
 Separation:  $\min \sum \|p_k - p_l\|^2$  (2)  
 Cluster validity index  $V_{xb} = \text{Compactness} / \text{Separation}$  (3)  
 where  $\| \cdot \|$  specifies the usual Euclidean norm,  $a_k$  is the  $k$ th data object,  $p_l$  is the  $l$ th clustering center . Compactness is calculated based on the distance from the data object to the cluster center where as separation is measured as the minimum distance between the center of the clusters.



**Figure2: Grid multidimensional K-medoids (GMK) method**

The grid clustering method detects the outliers, and provides natural and accurate clustering method for large dataset. The more the separation between clusters will be and the lesser the distance between objects within the same cluster will be, the more accurate the clustering will be. [12].

## 6. EXPERIMENTAL RESULTS

The accuracy of the proposed algorithm is compared with the existing one using ADULT Dataset [13]. The dataset contains 14 different classes and 32,561 instances. The attributes age with corresponding hours-per-week with all the instances is considered in order to find the appropriate cluster for the dataset. Table1 represents the snapshot of our experimental dataset.

**Table 1: Snapshot of adult data set**

Age	Hours-per-week
39	40
50	13
38	40
53	40
28	40
37	40

The accuracy of proposed GMK method and existing K-medoids method is depicted in the below table as table2.

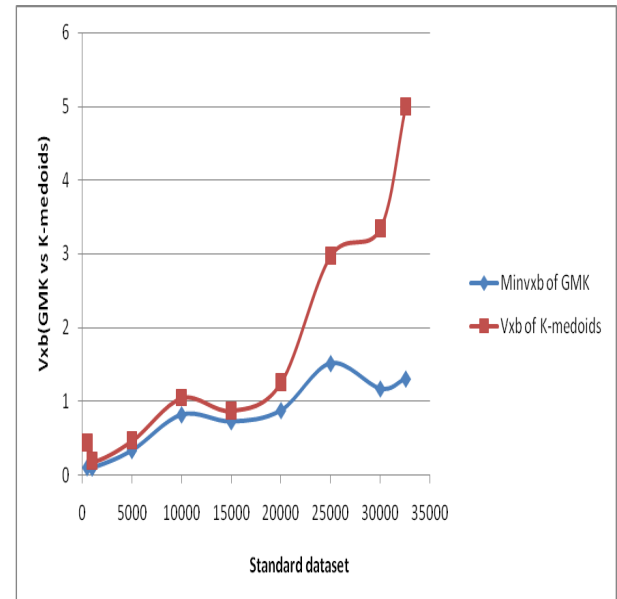
**Table 2: Accuracy of proposed GMK and conventional K-medoids.**

Dataset	Number of Cluster (As per GMK)	Value of m	MIN $V_{xb}$ of GMK	$V_{xb}$ of K-medoids
500	1	20	0.0992	0.443
1000	1	20	0.094	0.19
5000	2	20	0.331	0.4704
10000	4	20	0.818	1.049
15000	6	20	0.725	0.8763
20000	4	15	0.876	1.2575
25000	12	20	1.516	2.973
30000	37	5	1.172	3.341
32561	18	20	1.307	4.997

The cluster validity index of GMK is depicted as MIN  $V_{xb}$  where as the cluster validity index of K-medoids is depicted as  $V_{xb}$ . It is seen that for most cases if the value of m is kept as 20, more accuracy is found for the data set. It is also seen that for each and every case the value of cluster validity index is lower than the conventional K-medoids method which indicates that the proposed GMK method outperforms in terms of accuracy than the conventional K-medoids method.

In figure 3, which is applied on the two dimensional (2D) benchmark ADULT dataset, it is seen, that GMK has higher accuracy than the conventional K-medoids method as the cluster validity index is lesser. It implies that the objects in

GMK lie closer to each other than the conventional k-medoids method.



**Figure 3: Comparison of cluster validity index between the MIN  $V_{xb}$  of proposed GMK in terms of m and K-medoids method**

## 7. CONCLUSION AND FUTURE WORK

The intra cluster distance between objects in the proposed method is denser than the conventional K-Medoids clustering method. Also there is a tendency of gap between the merged clusters for large data set. Thus, the proposed GMK method has better accuracy than the conventional K-Medoids method. The proposed method is implemented for 2D data set and is expected to work better for higher dimensional data set. The use of principle component analysis could be used for the proposed GMK method which will provide better accuracy than the conventional K-Medoids method when the dimension is high.

## 8. REFERENCES

- [1] Han Jiawei and Kamber Micheline, 2006, "Data Mining Concepts and Techniques", second ed, China Machine Press.
- [2] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander, 2000, "Spatial data mining: database primitives, algorithms and efficient DBMS support", Data Mining and Knowledge Discovery, Kluwer Academic Publishers.
- [3] Cadez I., Smyth P. and Mannila H. 2001, "Probabilistic modeling of transactional data with applications to profiling, Visualization, and Prediction", In Proc of the 7th ACM SIGKDD, San Francisco, pp. 37-46.
- [4] Cooley R., Mobasher B. and Srivastava J, 1999 "Data preparation for mining world wide web browsing", Journal of Knowledge Information Systems, vol 1, pp 5-32
- [5] A. Ben-Dor and Z. Yakhini, 1999, "Clustering gene expression patterns" In Proc of the 3rd Annual

International Conference on Computational Molecular Biology (RECOMB 99), Lyon, France, pp11-14.

- [6] A.Jain, R. Dubes, 1988. “Algorithms for Clustering Data” Prentice-Hall, EnglewoodCliffs, NJ.
- [7] E. Koltach, 2001. “Clustering Algorithms for Spatial Databases: A Survey”, Department of Computer Science, University of Maryland.
- [8] W. Wang, J. Yang, and R. Muntz, 1997 “STING: a statistical information grid approach to spatial data mining”, In Proc of the 23rd VLDB Conference, Athens, Greece, pp.186-195.
- [9] R. Ng, and J. Han, 1994, “Efficient and effective clustering methods for spatial data mining” *In Proceedings of the 20th Conference on VLDB*, Santiago, Chile, pp.144-155.
- [10] Su Youli, Yi, Guohua Chen Liu, 2009, “GK-means: An Efficient K-means Clustering Algorithm Based On Grid”, School of Information Science and Engineering Lanzhou University, *In Proc. Of the International symposium on Computer network and multimedia Technology (CNMT)*, Wuhan, pp- 1 – 4.
- [11] [http://en.wikipedia.org/wiki/Flood\\_fill](http://en.wikipedia.org/wiki/Flood_fill)
- [12] Pardeshi and Bharat, Toshniwal Durga, “Improved K-Medoids Clustering Based on Cluster Validity Index and Object Density”, In Proc of IEEE 2<sup>nd</sup> International Advance Computing Conference, 2010, Indian Institute of Technology Roorkee, pp.379-384.
- [13] Zadrozny Bianca and Elkan. Charles, 2002. “Transforming classifier scores into accurate multiclass probability estimates”. *In Proc of the International Conference on Knowledge Discovery and Data Mining (KDD’02)*.