Available online at www.elixirpublishers.com (Elixir International Journal)

Computer Science and Engineering

Elixir Comp. Sci. & Engg. 71 (2014) 24818-24821

ISSN: 2229-712X

A Novel Approach for PAM Clustering Method

Faisal Bin Al Abid

Department of CSE, Institute of Science and Techno, affiliated with National University of Bangladesh.

ARTICLE INFO

Article history:
Received: 24 November 2013;
Received in revised form:
25 May 2014;
Accepted: 7 June 2014;

Keywords
Medoid, Grid,
ADULT Dataset,
Partitioning,
Time complexity,
Dense grid,
Outlier detection.

ABSTRACT

Existing and in recent times proposed clustering algorithms are studied and it is known that the k-means clustering method is mostly used for clustering of data due to its reduction of time complexity. But the foremost drawback of k-means algorithm is that it suffers from sensitivity of outliers which may deform the distribution of data owing to the significant values. The drawback of the k-means algorithm is resolved by k-medoids method where the novel approach uses user defined value for k. As a result, if the number of clusters is not chosen suitably, the accuracy will be minimized. Even, K-medoids algorithm does not scale well for huge data set. In order to overcome the above stated limitations, a new grid based clustering method is proposed, where time complexity of proposed algorithm is depending on the number of cells. Simulation results show that, the proposed approach has less time complexity and provides natural clustering method which scales well for large dataset.

Introduction

Data Mining is the method of non-trivial extraction of implicit, previously unknown, and potentially helpful information from data [1].Data mining tasks can be classified into two categories: descriptive and predictive. The illustration of the general properties of the data in the database is performed on descriptive data mining tasks i.e. descriptive task finds the human-interpretable patterns that express the data. Deduction on the existing data is performed in order to make predictions in the case of predictive mining tasks. In the midst of the descriptive data mining tasks clustering is one of the most important tasks. Representing the dataset by smaller amount of clusters misplaces certain fine details, but gains simplification [1]. Data modeling puts clustering in a chronological viewpoint rooted in mathematics, statistics, and numerical analysis. Clustering can be taken as a density assessment problem. From the perspective of machine learning, clusters correspond to hidden patterns, the examination for clusters is unsupervised learning, and the resulting system represents a data notion. From a practical perspective clustering plays a awe-inspiring role in data mining appliance 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 theme of vigorous investigation in several fields such as statistics, pattern recognition, and machine learning. There are huge datasets containing many attributes of diverse types. This kind of problems is added to clustering in data mining. This imposes sole computational rudiments on relevant clustering algorithms. 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.

Owing to the sensitivity of outlier in k-means, k-medoids clustering method is dealt with. 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 contrary, the disadvantages of k-medoids are:

(i) The user has to give the value of k.

(ii) It does not cope well for large data set.

In this paper, the main concentration was on eliminating these disadvantages using grid clustering GKmedoids approach. In this approach, first the grid was partitioned and the objects are put inside the grid. The neighboring grids are merged in 4 ways using Flood fill algorithm and the objects are put inside the grid. Each merged grid is considered as a cluster [6] and the GSA search algorithm is used in order to find the center [7]. 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 outliers are also detected based on Gravitational search algorithm. The global optimum was chosen as a minimization problem, that is, if the data points were closer together based on fitness function, then the neighboring cells were merged otherwise it would not be merged. Also a heuristic approach was made in order to find out different feasible parameters.

The rest of the paper is organized as follows: Section 2 presents the review literature of clustering. Section 3 and 4 describes the variants of k-medoids method and the conventional partitioning around medoids k-medoids method respectively. Section 5 presents the proposed Grid K-medoids (GKmedoids) method. Section 6 illustrates the experimental results using adult dataset. Finally, section 7 contains the concluding remarks.

Tele:
E-mail addresses: faisaliut42@yahoo.com

Review literature of clustering

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 [8]. 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 [9]. 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 [10].

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 result 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<<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.

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.

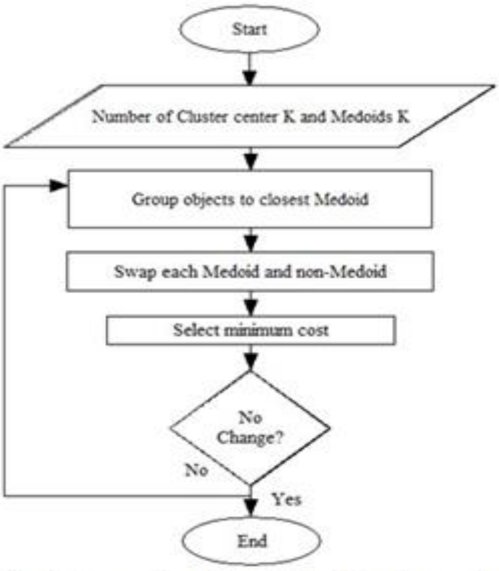

Start
Number of Cluster center K and Medoids K
Group objects to closest Medoid
Swap each Medoid and non-Medoid
Select minimum cost
No Change?
No
Yes
End

Fig 1: Conventional k-medoids clustering method

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 [11]. 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, the focus 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.

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) *o* 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.

Proposed GKmedoids clustering method

The concept of GK-means method was taken to partition the grid [12], according to that method. The algorithmic procedure of GKmedoids is subdivided into four main parts: Getting special grids, Detection of outlier, Cell merging, and Center selection of merged clusters. The algorithm is divided into sub-algorithms and the description of each and every part of the sub-algorithms is given below.

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)
        minX ⟵ point.getX()[0]
    end if
    if(point.getX()[1] < this.minY)
        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 $\sigma = \sqrt{\frac{N}{m}}$ 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 , σ_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.

Detection of outlier

Algorithm2:

Is Out(int w, int max)

fit(t) represent the fitness value of the point i at time t, and, worst(t) and best(t) are defined as follows (for a minimization problem):

fit_a (t) represent the fitness value of the agent a at time t, and, worst(t) and best(t) are defined as follows (for a minimization problem):

best (t)= min fit_b (t)
b€{1….N}

worst (t)= max fit_b (t)
b€{1….N}

The mass of the agent is determined by the following:

$M_a = m_a (t)/\sum m_b(t)$,
B={1….N}

where m_a (t)= fit_a (t)-worst(t)/ best (t)- worst (t)

```
If  Kbest <  fit_b (t)&& m_b(t),
Outlier: = cluster
Else
Cluster:=Merge(cells):
            end if
```

Description of algorithm2:

The mass of the data points is determined by the fitness function, the Kbest data points with the best fitness function and mass will be chosen in order to merge the cells , otherwise the cells won't be merged in order to form the cluster. This procedure of detection of the outliers is much more feasible than the method of determining outlier as in [6].

Cell merging

ALGORITHM3:

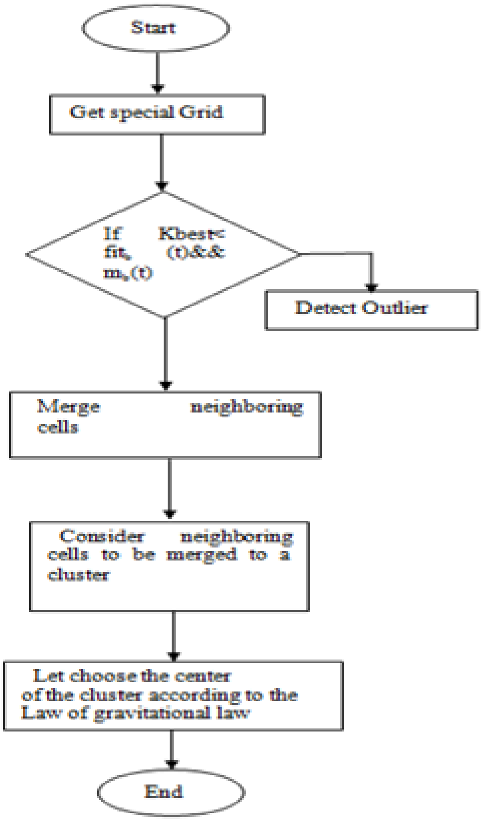
```
Merge ()
Cell Cluster []
Insert parent into Q
While (!Q. empty)
Cell Parent⟵Q. front ()
Q. pop ()
For all neighboring child of Parent
cluster. add (child)
set child visited
end while
```

Description of algorithm3:

The neighboring cells are merged by searching the adjacent four neighboring grids: Top, Bottom, Right and Left. The neighboring grids are merged using the flood fill algorithm [13].

Center selection of merged cluster

Flowchart: GKmedoids method


Start
Get special Grid
If Kbest< fit_b (t)&& m_b(t)
Detect Outlier
Merge neighboring cells
Consider neighboring cells to be merged to a cluster
Let choose the center of the cluster according to the Law of gravitational law
End

```
Udist (Point a, Point b)
ReturnMath.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]))
F_a(t)=∑rand_b F_ab (t), F_ab= G M_a M_b / Udist
Cluster (Center):= Max (F_ab )
```

Description of algorithm4:

The Euclidean distance between the data points are calculated and the force between the objects, that is the data points are calculated. The heavier mass will attract the other object that is, the heavier mass will have the least tendency to move but the most tendencies in order to attract the other objects. The heaviest mass is chosen as the center. The heaviest mass will attract the other objects; the intra cluster distance between the objects will be lesser in each and every cluster.

Experimental Results

We have evaluated the performance of our proposed algorithm using adult dataset [12]. The dataset contains 14 different classes and 32,561 instances. We have considered the attributes *age* with corresponding *hours-per-week* with all the instances in order to find same cluster of objects together.

Table 1: Snapshot of adult dataset

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

The output of the proposed GKmedoids method and conventional K-medoids method is performed on windows 7 platform with Pentium dual core processor 2.5 GHz, and with 2.00GB of RAM. While calculating the time the I/O time for reading the file was blocked and 10 averages of the output time where taken which were pretty much similar.

The output is depicted in the below table as table2.

Dataset	Number of Cluster (As per GKmedoids)	$\text{MIN}_{\text{time}}$of GKmedoids(in ms)	Time of K-mediods (in ms)
500	1	8	47
1000	2	14	110
5000	2	28.2	141
10000	7	49.7	265
15000	15	59.4	484
20000	29	79.4	875
25000	32	95	937
30000	33	99.9	1062
32561	27	107.2	1202

The time of GMK is depicted as MIN_{time} as it has minimized the time of conventional K-medoids method. The number of clusters is found from the Proposed GMK method and the same number of clusters is used to find out the time for the conventional K-medoids method. In figure 3, which is applied on the two dimensional (2D) dataset, it is seen, that GMK has much less time complexity than the K-medoids algorithm and the time decreases as the number of input increases. It is because in case of grid clustering the time is independent on the number of input but dependent on the number of cells in the grid.

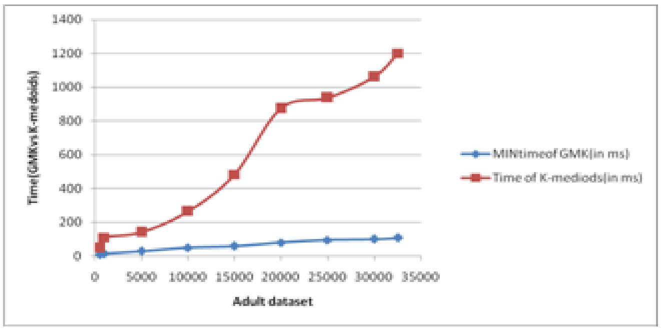


Fig 3: Proposed GMK has less time complexity than k medoids

From the figure above, it is seen that the more the number of data increase in the data set, the greater the time complexity is reduced for the grid clustering method.

Conclusions and future work

The proposed method is able to eliminate the outlier grids by considering using fitness function in gravitational search algorithm. The performance evaluation of the proposed method shows lower time complexity and faster processing than the existing conventional k-medoids method. Furthermore, the accuracy of the proposed grid K-medoids method is expected to be higher because of the intra-cluster similarity. The higher accuracy of the proposed GK method and using of different distance functions in order to find feasible cluster can be a good area of research in future. Furthermore, the time complexity of Grid -based PAM Clustering Method can be compared with the proposed method in this paper.

References

[1] Han Jiawei and Kamber Micheline, "Data Mining Concepts and Techniques", second ed, China Machine Press, 2006**.**

[2] M. Ester, A. Frommelt, H.-P. Kriegel, and J. Sander "Spatial data mining: database primitives, algorithms and efficient DBMS support." Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Volume 4, 2000,pp 193-216.

[3] Cadez I., Smyth P. and Mannila H. "Probabilistic modeling of transactional data with applications to profiling, Visualization, and Prediction", *In Proc of the7th ACM SIGKDD*, 2001, San Francisco, pp. 37-46.

[4] Cooley R., Mobasher B. and Srivastava J. "Data preparation for mining world wide web browsing", *Journal of Knowledge Information Systems*, vol 1, pp 5-32, 1999.

[5] A. Ben-Dor and Z.Yakhini "Clustering gene expression patterns" *In Proc of the 3rd Annual International Conference on Computational Molecular Biology (RECOMB 99)*, 1999, Lyon, France, pp11-14.

[6] Faisal Bin Al Abid, M.A. Mottalib," An Accurate Grid -based PAM Clustering Method for Large Dataset", *International Journal of Computer Applications (0975 – 8887)* ***Volume 41– No.21, March 2012***

[7] Esmat Rashedi, Hossein Nezamabadi-pour, Saeid Saryazdi, GSA: A Gravitational Search Algorithm, Information Sciences, 179 (2009) 2232–2248

[8] A. Jain, R. Dubes. "Algorithms for Clustering Data" Prentice-Hall, EnglewoodCliffs, NJ, 1988.

[9] E. Koltach. "Clustering Algorithms for Spatial Databases: A Survey", Department of Computer Science, University of Maryland, 2001. Available at: http://www.cs.umd.edu/~kolatch/papers/SpatialClustering.pdf.

[10] W. Wang, J. Yang, and R. Muntz, "STING: a statistical information grid approach to spatial data mining", *In Proc of the 23rd VLDB Conference*, 1997, Athens, Greece, pp.186-195.

[11] R. Ng, and J. Han, "Efficient and effective clustering methods for spatial data mining" *In Proceedings of the 20th Conference on VLDB*, 1994, Santiago, Chile, pp.144-155.

[12] Su Youli,Yi , Guohua Chen Liu, "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 ,2009,pp- 1 – 4.

[13] http://en.wikipedia.org/wiki/Flood_fill

[14] L. Blake and C. J. Merz. UCI repository of machine learning databases. Department of Information and Computer Sciences, University of California, Irvine, 1998. http://www, ies. uei. edu/~mlearn/MLRepos irony, html.