ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

DYNAMIC CLUSTRING ALGORITHM IN ASSORTED APPLICATIONS USING RANDOM FREEDOM FACTOR

Rubika Walia

M.Tech. Research Scholar

Computer Science and Engineering

M. M. University, Sadopur

Ambala, Haryana, India

Er. Neelam Oberoi

Assistant Professor

Computer Science and Engineering

M. M. University, Sadopur

Ambala, Haryana, India

ABSTRACT

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. It is one of the major tasks that is performed in Data Mining. Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. 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. Aside from the raw analysis step, it involves

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating.

INTRODUCTION

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

- 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.

APPLICATIONS

- 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.

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

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.

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

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

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 $k \leq n$. 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.
- 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
- Divisive Approach

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

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

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.

APPROACHES TO IMPROVE QUALITY OF HIERARCHICAL CLUSTERING

Here are the two approaches that are used to improve the quality of hierarchical clustering –

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

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

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

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.

DYNAMIC CLUSTERING

Dynamic clustering is a technique to find entries in your log similar to the current situation. Essentially, it is a K-nearest neighbor algorithm, and not actually clustering at all. Despite this misnomer, the term "Dynamic Clustering" has stuck with the Robocode community.

The idea is to record a "state" (or termed "situation") for each entry in your log. The state can contain any data that you deem valuable, such as lateral velocity, advancing velocity, or enemy distance. Save this along with your data. Then to use the data, you find a "distance" between current state and past states. Distance can be Euclidian ($\sqrt{(dist1 - dist2)^2 + (lat1 - lat2)^2 + \cdots}$) or another way, such as Manhattan distance ($|dist1 - dist2| + |lat1 - lat2| + \cdots$). Find some number of entries with the lowest distance, and use them for targeting, movement, or whatever you like.

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

The earliest method doing this was by iterating through the log and calculating the distance for each log entry. If you have a large log this is very slow. More recently kd-trees have been used. Corbos was the first one to mention them on the RoboWiki, which caught the interest of Chase-san and Simonton.

FUZZY CLUSTERING

In fuzzy clustering, every point has a degree of belonging to clusters, as in fuzzy logic, rather than belonging completely to just one cluster. Thus, points on the edge of a cluster, may be *in the cluster* to a lesser degree than points in the center of cluster. An overview and comparison of different fuzzy clustering algorithms is available.

Any point x has a set of coefficients giving the degree of being in the kth cluster $w_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster:

$$c_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}.$$

The degree of belonging, $w_k(x)$, is related inversely to the distance from x to the cluster center as calculated on the previous pass. It also depends on a parameter m that controls how much weight is given to the closest center. The fuzzy c-means algorithm is very similar to the k-means algorithm:

- Choose a number of clusters.
- Assign randomly to each point coefficients for being in the clusters.
- Repeat until the algorithm has converged (that is, the coefficients' change between two iterations is no more than ε, the given sensitivity threshold) :
    - Compute the centroid for each cluster, using the formula above.

ISSN (Online) : 2348 - 2001

International Refereed Journal of Reviews and Research

Volume 3 Issue 2 March 2015

International Manuscript ID : 23482001V3I2032015-05

(Approved and Registered with Govt. of India)

- For each point, compute its coefficients of being in the clusters, using the formula above.

The algorithm minimizes intra-cluster variance as well, but has the same problems as *k*-means; the minimum is a local minimum, and the results depend on the initial choice of weights.
Using a mixture of Gaussians along with the expectation-maximization algorithm is a more statistically formalized method which includes some of these ideas: partial membership in classes. Another algorithm closely related to Fuzzy C-Means is Soft K-means.

Fuzzy c-means has been a very important tool for image processing in clustering objects in an image. In the 70's, mathematicians introduced the spatial term into the FCM algorithm to improve the accuracy of clustering under noise.

REFERENCES

[1] URL - http://robowiki.net/wiki/Dynamic_Clustering

[2] Nock, R. and Nielsen, F. (2006) "On Weighting Clustering", IEEE Trans. on Pattern Analysis and Machine Intelligence, 28 (8), 1–13

[3] Bezdek, James C. (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. ISBN 0-306-40671-3.

[4] Ahmed, Mohamed N.; Yamany, Sameh M.; Mohamed, Nevin; Farag, Aly A.; Moriarty, Thomas (2002). "A Modified Fuzzy C-Means Algorithm for Bias Field Estimation and Segmentation of MRI Data" (PDF). IEEE Transactions on Medical Imaging 21 (3): 193–199. doi:10.1109/42.996338. PMID 11989844.