ISSN No. 0976-5697

Volume 8, No. 4, May 2017 (Special Issue)

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Study on Clustering Based Methods

Er. Paramvir Kaur Dhillon
CSE (Computer Science and Engineering)
SBBSU (Sant Baba Bhag Singh University)
Jalandhar, India
dparamvir8@gmail.com

Er. Amandeep Singh Walia
CSE (Computer Science and Engineering)
SBBSU (Sant Baba Bhag Singh University)
Jalandhar, India
er.amanwalia@hotmail.com

***Abstract*-- Data mining is a process of mining information from huge data sets and making it into a logical form for supplementary purpose. Clustering is an important step in data mining applications and data analysis. Clustering is a method of arranging objects with similar properties into a single group. Data mining is done by passing through various phases. The process of mining can be done by using two learning sets, supervised and unsupervised learning. The clustering is an unsupervised learning. It has following categorized partition-based method, density-based method and grid-based method, hierarchical-based method. A good clustering method will produce high superiority clusters with high intra-class similarity and vice-versa.**

***Keywords*- Data mining, clustering, classification of clustering, DBSCAN, Density Based clustering.**

I. Introduction

The purpose of the data mining technique is to mine information from a bulky data set and make over it into a reasonable form for supplementary purpose. Data mining is also known as the analysis step of the knowledge discovery in databases (KDD). Knowledge discovery means to “develop something new”. Data mining practice has the four main everyday jobs. These are Anomaly detection, Association, Classification, Clustering. Anomaly detection is the recognition of odd data records, that may be remarkable or data errors that involve further investigation. Association rule learning is the process to find the relationships between the variables. In this, relations are set up between the variables to create the new information that is needed for some purpose. Classification is the assignment of generalizing the known structure to apply to new data like in an e-mail process might attempt to categorize an e-mail as "legitimate" or as "spam". Clustering is a significant task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups (clusters). Cluster is an ordered list of data which have the familiar characteristics. Data mining is a multi-step process. In data mining data can be mined by passing through various phases [1].

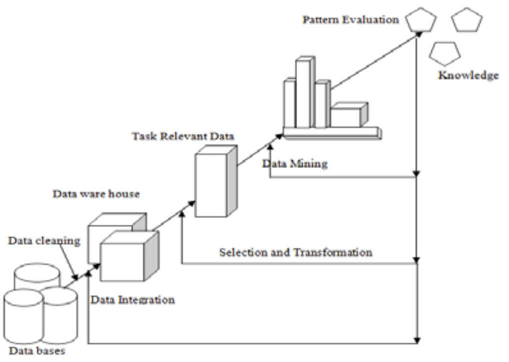


Figure 1: Data Mining Process [2]

These days Informational technology is mounting and databases created by organizations and companies like telecommunications, banking, marketing, transportation, manufacturing, and social networking sites etc. are becoming huge day by day. Knowledge discovery process is used to store this data in databases and efficiently access the interested or useful data from databases [2].

Knowledge discovery consist of following steps:

a) Data Cleaning: It is the step in which the process of detecting and removing of data which is not correct, irrelevant, containing missing values, duplicate values and noise that is dirty data from the database.

b) Data Integration: It is the step in which data from different sources is collected in one source to provide unified view of data.

c) Data Selection: It is the step in which data analysis is done in way that the selection of relevant data from databases.

d) Data Transformation: It is the step in which the data which is selected is reformed to correct form performing various operations like summary, aggregations, generalizations and normalized operations.

e) Data Mining: This is important technique in which intelligent operations are used to extract the useful pattern from the database.

f) Pattern Evaluation: It is the step in which the required pattern are evaluated from the given database.

CONFERENCE PAPERS
National Conference on Emerging Trends on Engineering & Technology (ETET-2017)
On 21st April 2017
University Inst. of Engg. & Tech. & University Inst. of Computer, SBBS University, Punjab (India)

g) Knowledge Representation: It is the step where whole process of output is visualized to user in the form of graphs, tables and graphs etc.

A. Classification of Data Mining System: According to following categories data mining system is classified:

a) According to Data source to be mined: Data mine system can be classified according to mined techniques used like spatial data, multimedia data ,time-series data etc.

b) According to Data models: Data mine systems may use many models like relational model, object oriented model and transactional models.

c) According to kind of Knowledge mined: Data mine system can be classified according to the type of knowledge is used like classification, prediction, cluster analysis and outlier analysis.

d) According to utilized Mining technique: Data mine system can be classified according to techniques used for data mining techniques like decision tree, neural network etc.

e) According to adapted applications: Data mine systems can be classified according to applications adapted like in finance, data mining system related to finance is used.

B. Major issues in Data Mining:

There are various data mining algorithms and techniques but there is enormous volume of data in world and there is continuous spike in the data, major issues that can be raised in data mining systems can be scalability and reliability of performance of data mining system [2].

Various performance issues are:

a) Effective, Efficient and Scalable data mining: in order to efficiently extract the useful knowledge from the large amount of databases, the technique of data mining which we are using should be effective, efficient and scalable, gives desired outputs in the desired time.

b) Parallel, Distributed and Incremental mining algorithms: The amount of data present in the databases is enormous and to maintain the complexity of data, data mining techniques helps to develop the parallel and distributed data mining algorithms. Data used in these algorithms is stored in different partitions and processed in parallel. The output generated from these partitions is combined to provide accurate results and this is quite tough job to mine data without any scratch.

C. Learning sets in data mining:

In Data Mining the two types of learning sets are used, they are supervised learning and unsupervised learning.

a) Supervised Learning: In supervised training, data includes together the input and the preferred results. It is the rapid and perfect technique. The accurate results are recognized and are given in inputs to the model through the learning procedure. Supervised models are neural network, Multilayer Perception and Decision trees [1].

b) Unsupervised Learning: The unsupervised model is not provided with the accurate results during the training. This can be used to cluster the input information in classes on the basis of their statistical properties only. Unsupervised models are for dissimilar types of clustering, distances and normalization, k-means, self-organizing maps [1].

II. CLUSTERING

Clustering is a crucial task in data analysis and data mining applications. It is the assignment of combination a set of objects so that objects in the identical group are more related to each other than to those in other groups. Cluster is defined as an ordered list of data items which have the familiar characteristics. Cluster analysis can be done by finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters. Clustering is an unsupervised learning process. A good clustering technique will produce high superiority clusters containing high intra-class similarity and low inter-class similarity. The superiority of a clustering result highly depends on equally the similarity measure used by the technique and its implementation. The superiority of a data clustering technique is also calculated by its ability to find out some or all of the hidden patterns in the data set. Similarity of a cluster can be expressed by the distance function. In data mining, there are some requirements for clustering the data [3]. These requirements are Scalability, Ability to deal with different types of attributes, Ability to handle dynamic data, discovery of clusters with arbitrary shape, Minimal requirements for domain knowledge to determine input parameters, Able to deal with noise and outliers, Insensitive to order of input records, High dimensionality, Incorporation of user-specified constraints, Interpretability and usability. The types of data sets that are processed for analysis of clustering are Interval scaled variables, Binary variables, and ordinal, Nominal and ratio variables, Variables of mixed types [4]. The five types of clusters are used in clustering. The clusters are divided into these types according to their characteristics. The types of clusters are Well-separated clusters, Center-based clusters

Contiguous clusters, Density-based clusters and Shared Property or Conceptual Clusters. Major applications of clustering are characterized by high dimensional data where each object is constituted of hundreds or thousands of attributes. Typical examples of high dimensional data can be found in the areas of computer vision applications, pattern recognition, and molecular biology [5]. The challenge in high dimensional is the curse of dimensionality faced by high dimensional data clustering algorithms, basically means the distance measures become gradually more worthless as the number of dimensions increases in the data set. Clustering has an extensive and prosperous record in a range of scientific fields in the vein of image segmentation, information retrieval and web data mining.

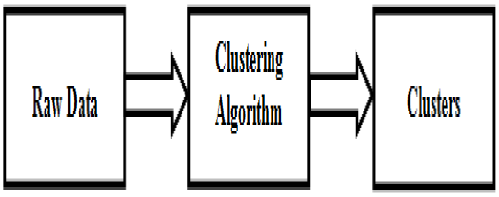


Figure 2: Stages of Clustering [3]

III. CLASSIFICATION OF CLUSTERING

In this section we will study the current and previous research work in spatial data mining and knowledge discovery. As we have discussed clustering plays a major role in understanding the usefulness of spatial data in real applications. So we will concentrate on meaning and methods of clustering in spatial data sets. Recent work carried out in the database community includes density-based methods, hierarchical methods, partition-based methods, grid-based methods, and constraint-based methods [6]. A brief idea of each and every method is given below with their positive and limited aspects.

A. Density-Based Methods: These kinds of methods consider clusters as dense region of objects that are different from lower dense regions in the data space. Density-based regions are highly accurate and can be useful in arbitrary shaped clusters, but choice of attributes and selection of clusters with algorithms is highly complex task. It has the feature to merge two clusters that are sufficiently close to each other. Density based sampling; DBSCAN (Density-Based Spatial Clustering of Applications with Noise), DENCLUE (DENsity CLUstEring), OPTICS (Ordering Points to Identify Clustering Structure) and so forth are instances of this method [7].

This method is our major discussion of this review paper so it will be discussed later in detail in the following sections.

B. Hierarchical Based Methods: Hierarchical based methods put the data in a tree-like structure. These types of clusters are classified into agglomerative & divisive hierarchical clustering, based on whether the decomposition is formed in a bottom-up or top-down manner. CURE (Clustering Using REpresentatives) BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies), CHAMELEON, and ORCLUS (arbitrary Oriented projected CLUSter generation) are the basic techniques of this category. These (Hierarchical Methods) are also able to recognize arbitrary shaped clusters and can handle outliers or noise excluding to some special conditions but this method is unable to work well for special characteristics of individual clusters and moreover it is time consuming for high dimensional data sets [7]. Hierarchical Clustering is classified as

1) Agglomerative Nesting: It is also known as AGNES. It is bottom-up approach. This method construct the tree of clusters i.e. nodes. The criteria used in this method for clustering the data is min distance, max distance, avg distance, center distance. The steps of this method are:

a) Initially all the objects are clusters i.e. leaf.
b) It recursively merges the nodes (clusters) that have the maximum similarity between them.
c) At the end of the process all the nodes belong to the same cluster i.e. known as the root of the tree structure.

2) Divisive Analysis: It is also known as DIANA. It is top-down approach. It is introduced in Kaufmann and Rousseeuw (1990). It is the inverse of the agglomerative method. Starting from the root node (cluster) step by step each node forms the cluster (leaf) on its own. It is implemented in statistical analysis packages [1], e.g., plus.

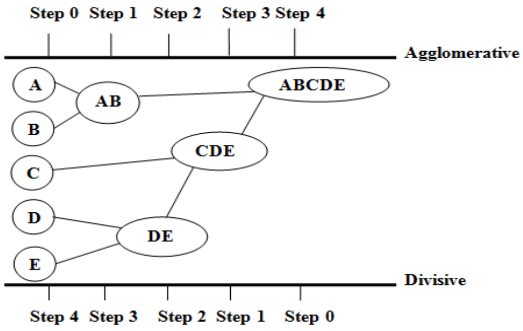


Figure 3: Representation of Agglomerative and Divisive [1]

Advantages of hierarchical clustering [8]

- Embedded flexibility with regard to the level of granularity.
- Ease of handling any forms of similarity or distance.
- Applicability to any attributes type.

Disadvantages of hierarchical clustering [8]

- Vagueness of termination criteria.

- Most hierarchical algorithm does not revisit once constructed clusters with the purpose of improvement.

C. Partitioning Methods: This method divides n objects, which we want to cluster, into k-partitions, where each partition represents a cluster and $k$is a given parameter. Such algorithms form the clusters to optimize an objective criterion similarity function such as distance as a major parameter. Partitioning methods cover the following five common algorithms: k-means, k-medoids, and CLARANS (Clustering Large Applications based upon RANdomized Search). Although partitioning methods are better in generation of clustering results by using k-mean, k-medoids are easier to implement but selection of $n$is random so no guarantee of quality of clustering and desired clusters is required in advance this is not more realistic. Dealing with outliers is also a major problem for these kinds of methods. A major flaw of this method is that it is not applicable for large data-sets or databases [7]. For a given k, the k-means algorithm consists of four steps:

a) Choose initial centroid at arbitrary.
b) Allocate each object to the cluster with the adjacent centroid.
c) Calculate each centroid as the mean of the objects assigned to it.
d) Reiterate previous 2 steps until no change.

D. Grid Based Methods: Grid based methods summarize the data space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. These types of methods are generally fast and independent of the number of the data objects. These dependent mainly on the number of the cells in each dimension in related space of generated outputs. [6].This method consists of the following well known algorithms STING (STatistical INformation Grid), wave cluster, CLIQUE (CLustering In QUEst), and STING+. These methods are able to automatically visualize subspaces of the highest dimensionality and are insensitive to the priority of records; moreover the accuracy of the clustering output may be degraded at certain points. Major applications of such kind of methods are military deployment, situation awareness, and so forth [7]. General steps followed in grid-based algorithm are:

1) Construct the grid structure, in other words division of the data space into a finite number of cells.
2) Finding the cell density for each of the cell
3) Sorting cells according to their densities.
4) Identification of the cluster centers.
5) Traversal of the neighborhood cells.

IV. DENSITY BASED CLUSTERING

Density based algorithms find the cluster according to the regions which grow with high density. These algorithms are known as one-scan algorithms. Basically, there are two approaches that may be used in density-based methods. The first approach, called the density-based connectivity clustering, pins density to a training data point. The algorithms that represent this behavior include DBSCAN and OPTICS. The second approach pins density to a point in the attribute space and is called Density Functions. This behavior is illustrated by the algorithm DENCLUE.

A. DBSCAN (Density-Based Spatial Clustering Of Application With Noise)

DBSCAN (Density Based Spatial Clustering of Applications with Noise) It is of Partitioned type clustering where more dense regions are considered as cluster and low dense regions are called noise [9].

Clusters are created on some of the basic criteria which are as follows:

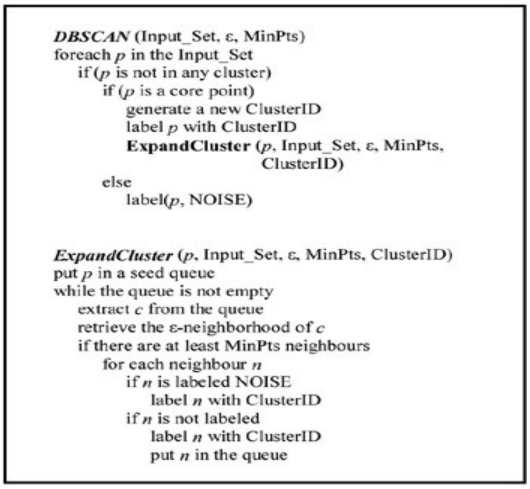
- Core: Core points lay in the interior of density-based clusters and should lie within Eps (radius or threshold value), MinPts (minimum no of points) which are user specified parameters.
- Border: Border point lies within the neighborhood of core point and many core points may share same border point.
- Noise: This point is neither a core point nor a border point..
- Directly Density Reachable: A point “r” is directly density reachable from s w.r.t Eps and MinPts if a point belongs to NEps(s) and |NEps (s)| >= MinPts.
- Density Reachable: A point “r” is density reachable from r point s wrt.Eps and MinPts if there is a sequence of points $r_1 \ldots r_n$, r_1 =s, r_n = s such that ri+1 is directly reachable from r_i.

Algorithm

Steps of algorithm of DBSCAN are as follows

1) Arbitrary select a point r.
2) Retrieve all points density-reachable from r w.r.t Eps and MinPts.
3) If r is a core point, cluster is formed.
4) If r is a border point, no points are density-reachable from r and DBSCAN visits the next point of the database.
5) Continue the process until all of the points have been processed.

The pseudo code of DBSCAN algorithm is shown in Figure.



```
DBSCAN (Input_Set, ε, MinPts)
foreach p in the Input_Set
    if (p is not in any cluster)
        if (p is a core point)
            generate a new ClusterID
            label p with ClusterID
            ExpandCluster (p, Input_Set, ε, MinPts,
                           ClusterID)
        else
            label(p, NOISE)


ExpandCluster (p, Input_Set, ε, MinPts, ClusterID)
put p in a seed queue
while the queue is not empty
    extract c from the queue
    retrieve the ε-neighborhood of c
    if there are at least MinPts neighbours
        for each neighbour n
            if n is labeled NOISE
                label n with ClusterID
            if n is not labeled
                label n with ClusterID
                put n in the queue
```

Figure 4: Pseudo code of DBSCAN [3]

ADVANTAGES OF DBSCAN:

- Unlike k-means, DBSCAN does not need one to specify the number of clusters in the data.
- It can discover arbitrarily shaped clusters. It can also find a cluster completely surrounded by a different cluster.
- DBSCAN works using two parameters only and it is not affected by the ordering of the points in the database.
- It is designed for use with databases that are capable of accelerating region queries, e.g. using an R* tree.
- The algorithm is robust to outliers.

DISADVANTAGES OF DBSCAN:

- DBSCAN is not completely deterministic.
- DBSCAN is incapable of clustering data sets well with large differences in densities.
- In order to choose a meaningful distance threshold ε, the data and scale must be well understood.

V. CONCLUSION

In this paper an up to date survey on clustering method and analysis of clustering in data mining is done. Partitioning clustering algorithm divides the data points into k partition, where each partition forms a cluster. Hierarchical clustering is a method of clustering which splits the identical dataset by making a hierarchy of clusters. Density based method discovers the cluster based on the regions which expands with high density. It is categorized under the one-scan algorithms. Grid Density based method use dense grids and multiresolution grid data structure to form clusters. Its main advantage is the fastest processing time.

REFERENCES

[1] A.K. Mann and N. Kaur, “Review Paper on Clustering Techniques”, Global Journal of Computer Science and Technology, Software and Data Engineering (0975-4350), Volume 13 Issue 5, pp 43-47, 2013.

[2] H.J. Jiawei, M. Kamber, “Data Mining: Concepts and Techniques (3rd ed.)”, Morgan Kaufmann publication, San Francisco, pp. 26-39, 2012.

[3] H. Shah, K. Napanda, L. D’mello, “ Density Based Algorithms”, in: IJCSE International Journal of Computer Sciences and Engineering, Volume 3, pp. 54-57, 2015.

[4] J. Daxin, C. Tang and A. hang “Cluster Analysis for Gene Expression Data”, A Survey, IEEE Transactions on Knowledge and Data Engineering, Volume 16, Issue 11, pp. 1370-1386, 2004.

[5] S. Jahirabadkar and P. Kulkarni “Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms”, International Journal of Computer Applications, Volume 63, pp. 29-35, 2013.

[6] M. Hemalatha, M. Naga, and N. Saranya, “A recent survey of knowledge discovery in spatial data mining,” International Journal of Computer Science, volume 8, no. 3, article 2, 2011.

[7] A. Sharma, R.K. Gupta, A. Tiwari, “Improved density based spatial clustering of application of noise clustering algorithm for knowledge discovery in spatial data” Hindawi Publishing corporation mathematical problems in engineering , pp.1-9, 2016.

[8] P. Rai and S. Singh, “A Survey of Clustering Techniques”, International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, pp. 1-5, 2010.

[9] P. B Nagpal and P. A Mann, “Comparative Study of Density based clustering Algorithm”, International journal of computer Application volume 27- No 11, pp. 44-47, 2011.

[10] V.S Ware, H.N Bharathi, “Study of Density based Algorithms”, International Journal of Computer Applications, Volume 69– No.26, 2013.

[11] K. Khan and S.R Rehman “DBSCAN: Past, present and future”. Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the. IEEE, pp 232-239, 2014.

[12] D.H. Widyantoro, T.R. Ioerger, J. Yen, “An incremental approach to building a cluster hierarchy”, ICDM Proceedings IEEE International Conference on Data Mining, pp. 705–708, 2002.

[13] S.A.L. Mary, K.R.S. Kumar, “A density based dynamic data clustering algorithm based on incremental dataset”, J. Computer Sci. Volume 8 (5), pp. 656–664, 2012.