Overview of Different Data Clustering Algorithms for Static and Dynamic Data Sets

Johnsymol Joy

Lecturer
Saintgits College Of Applied Sciences, Pathamuttam,Kerala
MG University

Abstract

Data mining is the process of extracting meaningful information from a large set of data. Data clustering is one of the major techniques used in data mining. These techniques will group related data in to identical groups. Data clustering is an unsupervised data analysis and data mining technique; it generates meaningful views from an inherent structure of data. Hundreds of clustering algorithms have been developed by researchers from a number of different scientific disciplines. Data may be static or dynamic. This paper focussed on different clustering algorithms for static and dynamic datasets.

Keywords:

Data mining, data clustering, data stream, Bayesian classifier, decision tree, Pattern mining etc

I. INTRODUCTION

Data mining is the process of extracting meaningful information from a huge set of data. Nowadays it is a vast growing field of computer science and information technology. These techniques can be used in various fields like banking, education, researches etc. Some important concepts of data mining include classification, data clustering, pattern mining etc. Classification means classifying data in to several groups based on some criteria. Such groupings are also termed as supervised learning, because we have a classifier and grouping is done based on that classifier. With the help of classification we can derive new rules and we can use that rules for analysing a new dataset. For example, consider a bank database; assume it may contain following fields like customer names, age, income etc. By analysing already existing data, it is possible to derive a new rule for checking a newly arrived customer may be eligible for a loan or not. Different classification techniques are Bayesian classifier, decision tree induction etc.

Pattern mining means using or developing data mining algorithms to find out interesting, unpredicted and valuable patterns in databases. It can be applied on various types of data such as transaction databases, sequence databases, streams, strings, spatial data, graphs etc.

The most popular algorithm for pattern mining is Apriori (1993). It is designed to be applied on a transaction database to learn patterns in transactions made by customers in stores. But it can also be applied in several other applications. Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Help users understand the natural grouping or structure in a data set. Used either as a stand-alone tool to get insight into data distribution or as a pre-processing step for other algorithms. Data collection may be static or dynamic.

II. DATA CLUSTERING APPROACHES FOR STATIC DATASETS

Data clustering approaches can be classified as Partition Clustering, Hierarchical Clustering, Density-based Clustering, Grid-based Clustering [1] etc.

A. Partitioning Methods

Partitioning method split entire data sets in to k-partitions and iteratively it finds k number of clusters. K-means [3] and k-medoid [7] are the two major examples of partition method clustering.

1). K-MEAN ALGORITHM

- Partitions data points into K clusters – K is predefined.
- Identify the initial cluster centers called centroids.
- Iteration until no change.
  - For each data points X_i, Calculate the distances between X_i and the K centroids.
  - (Re)assign X_i to the cluster whose centroid is the closest to X_i.
  - Update the cluster centroids based on current assignment.

K-mean have several advantages, some of them are it is easy to implement and faster than hierarchical method, if number of cluster K is small. It produces higher number of clusters. Their demerit includes difficulty in predicting the number of clusters. Initial seeds and the order of data have strong impacts on the final results. Rescaling dataset will entirely change the results.

B. Hierarchical Methods

The hierarchical method may be top-down or bottom- up. Here just focus on the basic bottom-up approach. Consider the case of N set of items. First assign each item to a cluster or we can say that initially it assumes each item as a cluster. Then finds the closest pair of clusters and merge them in to a single cluster. Then compute the distance between new cluster and the old clusters and repeat this until all items are clustered into a single cluster of size N. The most important advantages of hierarchical clustering [8] are it always output a structure that is more informative than k-means output and also easy to implement. Demerits are it is not possible to undo a previous step, once the instances have been assigned to a cluster they can no longer be moved around. It is not suitable for large datasets. It is very sensitive to outliers. The order of data can have a strong impact on final results.

C. Density Based Clustering

Density-based clustering algorithms [2] cluster data based on some connectivity and density functions. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [12] is one of the most widely used density based algorithm. It uses the notion of density reachability and density connectivity. A point "p" is said to be density reachable from a point "q" if point "p" is within ε distance from point "q" and "q" has sufficient number of points in its neighbours who are within distance ε. A point "p" and "q" are said to be density connected if there exist a point "r" which has sufficient number of points in its neighbours and both the points "p" and "q" are within the ε distance. Start with an arbitrary point and with the help of density reachability and density connectivity functions, it tries to expand and form cluster around itself.

Advantages includes, there is no need of prior requirement of number of clusters. It is capable of finding noise or outliers, while clustering also able to find arbitrarily shaped and arbitrarily sized clusters. It does not work properly for neck type and high dimensional dataset.

D. Grid Based Clustering

Grid based clustering quantize space into a finite number of cells that form a gri have a set of records and we want to cluster with respect to two attributes, then, we divide the related space (plane), into a grid structure and then we find the clusters. CLIQUE (CLustering In QUest) and STING (STatistical Information Grid) are the major types of this approach. The steps of CLIQUE are Partition the data space and find the number of points that lie inside each cell of the partition. Then identify the subspaces that contain clusters using the Apriori principle. Then identify clusters, for that determines dense units and connected dense units in all subspaces of interests. Then generate minimal description for the clusters, for that determines maximal and minimal regions that cover a cluster of connected dense units for each cluster.

Merits include its ability to automatically find subspaces of the highest dimensionality. There is no impact on the order of data. It scales linearly with the size of input and has good scalability as the number of dimensions in the data increases. Demerit is the accuracy of the clustering result may be degraded at the expense of simplicity of the method

III. DATA CLUSTERING APPROACHES FOR DATA STREAMS

Previous discussed algorithms works on static data sets. However, nowadays the need for processing data streams is an essential thing. We want to process sensor data, telecommunication operations, banking and stock- market applications; e-commerce etc and we want to analyze continuously arriving data streams. The growth of volume of live data and lack of data storage capacity will direct us to the dynamic processing of data and extracting knowledge. In this way data have been considered as a stream of data which come in from one side and exit from another side so we aren't able to visit data for the second time.

A) Stream

The STREAM [13[11] framework is based on the k-medians clustering methodology. The central part is to split the stream into chunks, each of which is of convenient size and fits into main memory. Thus, for the original data stream D, we divide it into chunks D1 ...Dr ..., each of which contains at most m data points. The value of m is defined on the basis of a pre-defined memory account. Since each chunk fits in main memory, a variety of more complex clustering algorithms can be used for each chunk. The methods can use a diversity of different k-medians style algorithms for this purpose.

B) HDC-Stream

HDC-Stream [15] is a hybrid density-based clustering algorithm for evolving data streams. It has online and offline components. For a data stream, at each timestamp, the online component of HDC-Stream continuously reads a new data record and either adds it to an existing minicluster or maps it to the grid. In pruning time, HDC-Stream periodically removes real outliers. The offline component generates the final clusters on demand by the user.

C) ODAC

Online divisive agglomerative clustering (ODAC) is a time series data stream clustering

technique [10]. This algorithm is capable to handle concept drift using both agglomerative and divisive hierarchical methods. It uses a top-down strategy to create a tree-like hierarchy of clusters. A correlation-based dissimilarity measure (splitting criterion) is utilized to split each node, followed by the use of agglomerative strategy to increase the detection of concept drift among the time series data.

D) Denstream

DenStream[6][14] is a density microclustering algorithm for developing data stream. The algorithm extends the microcluster [10] concept and introduces the outlier and potential microclusters to distinguish between outliers and the real data. It has online and offline phases. In the online phase, the microclusters are created and in the offline phase, there will be the process of macroclustering, that is clustering of microclusters .

IV. CONCLUSION

Data clustering is an important area in data mining. While clustering, we should handle both static and dynamic data sets. There are lot of algorithms available for clustering static datasets and dynamic data streams. This paper contains finding outs of the study on some of those algorithms, mainly its advantages and disadvantages. Data streams are continuously generated over time. In next paper, will focuses more on data clustering algorithms for data streams.

REFERENCES

[1] R.Xu and D. Wunsch, “Survey of Clustering Algorithms,” IEEE Transactions on Neural Networks, vol. 16, no. 3, pp. 645–678, May 2005.[Online]. Available:http:/dx.doi.org/10.1109/TNN.2005.845141

[2] M.Ester, H.-P. Kriegel, J. Sander, and X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise,” in Proc. of 2nd International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.

[3] D.Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms.

[4] A.Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognition, vol. 36, no. 2, pp. 451 – 461, 2003.

[5] J.a. Gama, P. P. Rodrigues, and L. Lopes, “Clustering distributed sensor data streams using local processing and reduced communication,” Intell. Data Anal., vol. 15, pp. 3– 28, Jan. 2011.

[6] A.AminiandT.Y.Wah ,“Density micro-clustering algorithms on data streams: a review,” in Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS ’11),pp.410–414,HongKong,March2011.

[7] F.Gullo, G. Ponti, and A. Tagarelli, “Clustering uncertain data via k-medoids,” in Proceedings of the 2Nd International Conference. Available: http://dx.doi.org/10.1007/978-3-540-87993-019.

[8] F.Gullo, G. Ponti, A. Tagarelli, and S. Greco, “A hierarchical algorithm for clustering uncertain data via an information-theoretic approach,” in Data Mining, 2008. ICDM’08. Eighth IEEE International Conference on. IEEE,2008,pp.821–826.

[9] M.Charikar, C. Chekuri, T. Feder, and R. Motwani, “Incremental clustering and dynamic information retrieval,” in Proceedings of the twenty-ninth annual ACM symposium on Theory of computing. ACM, 1997, pp. 626–635.

[10] P.P. Rodrigues, J. Gama, and J. P. Pedroso, "ODAC: Hierarchical Clustering of Time Series Data Streams," in SDM,2006.

[11] K.Udommanetanakit, T. Rakthanmanon, and K. I Waiyamai, "E-stream: Evolution-based technique for stream clustering," in Advanced Data Mining and Applications, ed: Springer,2007,pp.605-615.

[12] W.-K. Loh and Y.-H. Park, "A Survey on Density-Based Clustering Algorithms," in Ubiquitous Information Technologies and Applications, ed: Springer, 2014, pp. 775-780.

[13] S.Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O’Callaghan, “Clustering data streams: Theory and practice,” IEEE Trans. on Knowl. and Data Eng., vol. 15, no. 3, pp. 515–528, Mar. 2003. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2003.1198387

[14] A.AminiandT.Y.Wah,“Densitymicro-clustering algorithms on data streams: a review,” in Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS ’11),pp.410–414,HongKong,March2011.

[15] AminehAmini,HadiSaboohi,TehYingWah,andTututHeraw an,“ A Fast Density-Based Clustering Algorithm for Real-Time Internet of Things Stream”, Hindawi Publishing Corporation e Scientific World Journal Volume 2014, Article ID 926020, 11 pages http://dx.doi.org/10.1155/2014/926020.