International Journal of Engineering & Technology, 7 (4.6) (2018) 214-216

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Conceptual ClusteringAnalysis in Data Mining: A Study

K.Nikhila[1], P.Manvitha[2]

[1] *Asst.Professor,dept of IT,Vardhaman College of Engineering,Hyd.*
[2] *Asst.Professor,dept of IT, Vardhaman College of Engineering,Hyd.*
**Corresponding author E-mail: nikhilamtech@gmail.com*

Abstract

Clustering on unsupervised learning handles with instances, which are not classified already and not having class attribute with them. Applying algorithms is to find useful but items on unknown classes. Approach of unsupervised learning is about instances are automatically making into meaningful groups basing on its similarity. This paper we study about the basic clustering methods in data mining on unsupervised learning such as ensembles distributed clustering and its algorithms.

Keywords: *Clustering; Data Mining; Density-based; Hierarchical; k-means; Dendrogram.*

1. Introduction

Clustering is the main task of groping **Data mining** and common method for data analysis used in numerous fields next to bio-informatics, data compression, image analysis, machine learning, information retrieval, and computer graphics**.** Along with clustering other names with similar meanings are numerical taxonomy, classification, botryology, a Greek word βότρυς, in English "grape" **Clustering or cluster analysis[1]** is the process of physical or theoretical objects into module of similar objects or the task of similar objects in a cluster, making a set of objects, to other clusters. It can be achieved by various algorithms and methods that differ in notion and efficiency. Some clusters include with distances between members, particular empirical distribution. Cluster analysis is not an automatic method, but an iterative process of discovering knowledge. Optimizing multi-objective interaction includes trial and error.

2. Partitioning methods

The centroid and the medoids are the two important sub-categories in partitioning methods. The centroid algorithms are using gravity center and the medoid algorithms are using nearest to the gravity centre for the instances on each cluster. The most popular method is Lloyd's algorithm, often just referred as K-means algorithm, which comes under centroid; all points in a given cluster are nearest to the centre point, when data set divided into K partitions. Finding local optimum and executes multiple times restricting K-medoids with variable random initializations choosing a fuzzy cluster assignment, fuzzy C-means. Drawback of this algorithm is the value of 'k' must be specified in advance. Declaring k value in advance results incorrect borders of clusters. Some theoretical wise interesting properties of k-means algorithm are: i) **[1]Voronoi diagram**: data space partitioned into structure. ii) Just as in machine learning, concept wise similar to the closest to neighbor classification.iii) similar to the model based clustering variation along with Lloyd's algorithm. Expectation maximum algorithm is shortly as EM .it helps in finding maximum likelihood parameters. Direct equation solving will not be occurred in statistical model. Unknown parameters, latent variables and known data observations are included in these models.

Description:

Statistical Model generates, observed data of set X, unobserved latent data /missing values set as Z, Θ as a vector of unknown parameters with a likelihood function.

$$L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

Maximizing the marginal likelihood of the observed data will be calculated by the MLE (Maximum Likelihood Estimate) for the unknown parameters.

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X}|\boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) d\mathbf{Z}$$

Where, intractable quantity occurs often.There are two steps to implement, the EM algorithm to find MLE of the marginal likelihood are :

E step / Expectation step : under the present estimate X of parameters $\boldsymbol{\theta}^{(t)}$ can be calculated from the expected log likelihood function w.r.t the conditional distribution Z. Unknown parameters of a vector with a likelihood function observed data marginal likelihood can be maximized by calculating unknown parameters MLE [2].

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{(t)}} [\log L(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z})]$$

M step / Maximization step: Parameters can be found by maximizing the quantity.

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

Larger data sets processed in efficient manner, local optimum terminates frequently, spherical shaped clusters, noise sensitivity are the important properties of k-means algorithm. This is termed as a batch method as it needs all the data must be available in advance. In this limitation variants of k-means clustering will be attained. Selecting proper centroid initialization is the key of basic procedure. Recent partitioning clusters which are disjoint are k-modes algorithm. Higher confidence assigned to larger valued membership function, in every cluster for each pattern where associated in a fuzzy clustering extension notation.

In[4], The k-modes algorithm is a current partitioning algorithm, which uses matching coefficient in simple which deals categorical attributes. The k-prototypes algorithm, through dissimilarity is defined, along with the k-means and k-modes algorithms to evaluate for clustering instances explained by mixed attributes. Data clusters in ellipse-shaped are new generalization of conventional k-means without dead-unit, the problem with a ball-shaped also performs exact clustering without determining the cluster number in advance. This generates partitions using traditional clustering methods. Each related to one value which can be higher confidence to cluster of pattern in assignment.FCM (fuzzy C-means) algorithm is a widely used based on k-means and finding the characteristic point in every cluster. This can be considered as the cluster center and for each instance it grades the membership in the clusters. Further developed clustering algorithms are based on the EM. Probability model with parameters explains a particular clusters probability. In this it starts with initial guessing for mixture model parameters. These values are calculated cluster probabilities in every instance process repeats when the parameters are re estimated uses these probabilities. Main drawback is expense and over fitting in this algorithm. Two reasons that causes this situation are specifying large no. of clusters and also too many parameters are having probability distribution. Solution for this is to adopt complete Bayesian approach, where prior probability distribution in each parameter.

3. Hierarchical clustering

Cluster analysis seeking to construct hierarchy of clusters called as hierarchical clustering or hierarchical cluster analysis shortly it is called as HCA. That means grouping objects of data into a tree clusters. This is also called as connectivity based clustering Two types of hierarchical clusters are agglomerative and divisive.

- **Agglomerative**: Every observation starts in its own cluster when pairs are merged as it moves up the hierarchy or desired number of clusters is satisfied till termination condition attains. It is a "bottom up" approach.
- **Divisive**: In [3], Every observation starts in one cluster and splits up into smaller cluster in a "top down" and recursive approach will be taken place as it moves down the hierarchy.

Generally, the merges and splits are resolved in a greedy manner and are known as a quick termination. But as soon as splitting and merging are done they show inability to when adjustments are performed. In fig1,The hierarchical clustering results are represented in a dendrogram.HAC requires memory $\mathcal{O}(n^2)$, time complexity $\mathcal{O}(n^3)$ even medium data sets works too slowly. SLINK is abbreviated as a single –linkage and CLINK is abbreviated as a complete linkage clustering for efficient optimal agglomerative methods[5].

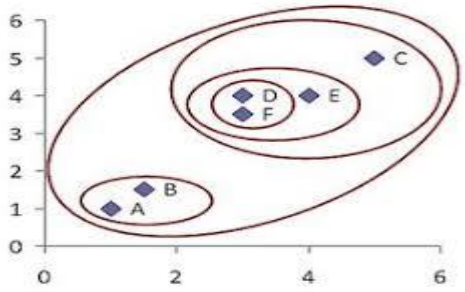

Fig 1: Hierarchical clusters represents on graph with objects

Along with this other advantages are : Numbers of clusters are not required in prior, calculate hierarchy of clusters completely, and best visualization results are involved into methods and later "flat "partition can derivates.

At every step which cluster should be joined or splitted will be decided "locally" in hierarchical clustering techniques. Proximity between two clusters defined by three definitions is: single-link, complete-link and average-link.in single-link, likeliness between any two clusters is the likeliness between two same instances, in every cluster one can resemble. It is good at non-elliptical shapes, but sensitive at noise and outliers.

In [4], complete-link, every cluster has one likeliness between two unequal instances. It is less sensitive to noise and outliers, but can split big clusters. Problematic is with convex shapes. Compromise between two clusters in average-link. Hierarchical clustering algorithms are: BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), CURE (Clustering Using Representatives) and CHAMELEON BIRCH uses CF-tree which is a hierarchical data structure used for partitioning the data which are incoming in dynamic and incremental process. It uses iterative relocation. Other two gives careful analysis "linkages" of object in every hierarchical partitioning.

4. Density-based clustering

In [9], a region finding clusters based on data point's density ids done by Density-based clustering algorithms. In a given radius (Eps) neighborhood for every instance to a cluster should contain minimum instances is the main key for this algorithm. In fig2, DBSCAN is most important method in density based clustering algorithms. It is abbreviated as Density-Based spatial Clustering of Applications with Noise. In 1996 it is proposed by Martin Ester, Hans-peter kriegel, Sander and Xiaowei Xu. In fig3, this data points separates in three classes they are: core points, border points and noise points. In core points, points are inside the cluster, if there are enough to its neighborhood. Border point is not within the border but falls to the neighborhood core point. Noise point is neither a core nor border point.

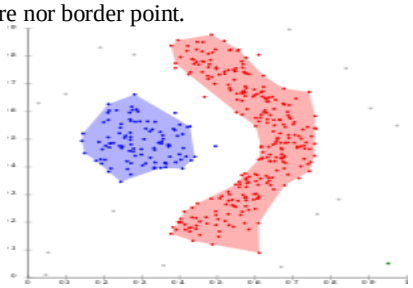

Fig 2: DBSCAN with Density-based clustering

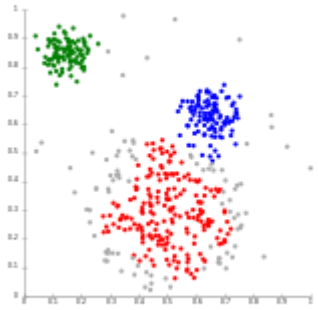

Fig 3: Assuming clusters of same density, and might have problems departing nearest clusters by DBSCAN[6].

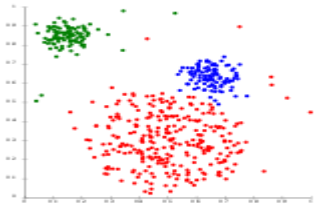

Fig4: DBSCAN variant is OPTICS which handles random densities in a better way.

In [7], recent studies by Clarans and Birch, processing Big data to increase the performance. Methods on pre-clustering development like canopy clustering. In fig4, these huge data sets processes efficiently but existing clusters are merely pre partitioned which are slower methods like k-means clustering along with seed based clustering. Organizing and analyzing data in high dimension is a curse dimensionality is the failure for existing methods, which leads to new clustering algorithms, focusing on subspace and co-relational clustering. Incremental version for DBSCAN are GDBSCAN (Generalizing the density-based algorithm),PDBSCAN(parallel version of DBSCAN),DBCLASD(Distribution Based Clustering of Large Spatial Data sets).OPTICS is one of the new algorithm ,which is versatile for interactive cluster analysis .It orders the data sets showing density based cluster structure. DENCLUE is another density based algorithm. Here the given radius ε to its neighborhood with minimum number of objects μ i.e. cardinality increases its threshold had some basic definitions

Definition 1 Core Object

In [6], a set D of object o is called core object along with ε and μ, if $|N\varepsilon\ (o)| \geq \mu$, where Nε is the subset of D contains in ε-neighborhood.

Definition 2 Directly Density-Reachable

In a set D from object 'o', object 'p' is directly density-reachable, when o is the core object p Є Nε (o).only from core objects other can be directly density- reachable.

Definition 3 Density-Reachable, Density-connected

In [8], if there are chain objects $P_1,P_2,\ldots Pn$ where $P_1=0,Pn=P$ such that pi Є D and from p_i density-reachable to p_{i+1}.In this density connectivity is a symmetric relation and density-reachability is a non-symmetric.

5. Conclusion

Generally, decomposing many levels of partitioning in data set representing by dendrogram in hierarchical clustering algorithm which is more effective. Dendrograms are expensive for big data analysis.

For better results and efficient purpose, instead of using individual algorithms better to use combinational to reduce drawbacks for particular algorithms. Combining multiple combinational algorithms in early stage needed expansion, along with that impact of coordinated sub sampling methods for object quality and efficiency is also needed. Here question is what type of clustering algorithms has to be combined to achieve better results and also reusing the knowledge.

References

[1] S.B. KOTSIANTIS, P. E. PINTELAS "Recent Advances in Clustering: A Brief Survey Department of Mathematics" University of Patras Educational Software Development Laboratory Hellas.

[2] Ankerst M., Breunig M., Kriegel H., Sander J., OPTICS: "Identify the Clustering points in its structure," Proc. ACM SIGMOD'99 Int. Conf. on Management of Data.

[3] H. Ayad and M. Kamel. "Basing on sharing nearest neighbor finding natural clusters using multi-clusters combination". In Multiple Classier Systems: Fourth International Workshop, MCS 2003, Guildford, Surrey, UK

[4] M. Breunig, H.-P. Kriegel, R. Ng, and J. Sander. Lof: "Identifying density-based local outliers." In Proc. of SIGMOD'2000, pages 93–104, 2000.

[5] Cheeseman P. & Stutz J., (1996), Bayesian Classification (AutoClass): Theory and Results, In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smith, and R. Uthurusamy, editors, "Advances in Knowledge Discovery and Data Mining."

[6] Yiu-Ming Cheung, k*-Means: "A new generalized k-means clustering algorithm, Pattern Recognition" Letters 24 (2003).

[7] Hans-Peter Kriegel, Martin Pfeifle ,"Hierarchical Density-Based Clustering of Unsupervised Data" Institute for Computer Science University of Munich, Germany.

[8] Hans-Peter Kriegel Martin Pfeifle "Density-Based Clustering of Uncertain Data." University of Munich, Germany University of Munich, Germany Institute for Computer Science Institute for Computer Science.

[9] Clustering of time series data—a survey T. Warren Liao∗Industrial & Manufacturing Systems Engineering Department, Louisiana State University, 3128 CEBA, Baton Rouge, LA 70803, USA.