Clustering approaches used for object oriented scenario

Sachin Saxena
Computer science & Engineering
Invertis University Bareilly
sachinrajsaxena2@gmail.com

Anil Pandey
Computer science & Engineering
Invertis University Bareilly
anil.p@invertis.org

Abstract

Object is a real entity and object having properties life reusable, maintainable and extendable. This paper introduces different clustering approach used in object oriented system. Clustering is a method used for dividing object into their related groups. In clustering process, grouping objects into cluster such that object from the same cluster are similar and object of different cluster are dissimilar. During the last five years research on and with the clustering algorithms has reached a very promising state. In this paper a brief survey on clustering algorithm is describe with the help of real based scenario.

Keywords: Clusters, objects

1. Introduction

Object-oriented refers to a programming language, system or software methodology that is built on the concepts of logical objects. It works through the creation, utilization and manipulation of reusable objects to perform a specific task, process or objective. Object oriented is a computer science concept that has been widely implemented, specifically in programming languages and applications/software. The object-oriented technique is different from conventional programming, which focuses on functions/behaviours, while object-oriented works on the interactions of one or more objects.

C clustering is the grouping of a particular set of objects based on their characteristics, aggregating them according to their similarities. Regarding to data mining, this methodology partitions the data implementing a specific join algorithm, most suitable for the desired information analysis.

This clustering analysis allows an object not to be part of a cluster, or strictly belong to it, calling this type of grouping hard partitioning. In the other hand, soft partitioning states that every object belongs to a cluster in a determined degree. More specific divisions can be possible to create like objects belonging to multiple clusters, to force an object to participate in only one cluster or even construct hierarchical trees on group relationships.

Object-oriented DBMS are still a relatively new area, with many unanswered questions as to their performance. Objects can be clustered on disk so that when accessing one object in a cluster, all of the objects in that cluster are brought into main memory. Thus when accessing additional objects in the cluster, it is then a main memory operation rather than a disk operation. It is widely acknowledged that a good object clustering is critical to the performance of OODB. Clustering means storing related objects close together on secondary storage so that when one object is accessed from disk, all its related objects are also brought into memory.

2. Different clustering approaches

According to Kedar B. Sawant [1] the Clustering is a well known data mining technique which is used to group together data items based on similarity property. They describe existing methods for selecting the number of clusters as well as selecting initial centroid points, and also give method for selecting the initial centroid points and the modified K-mean algorithm which will reduce the number of iterations and improves the elapsed time. An overview of the existing methods of choosing the value of K i.e. the number of clusters along with new method to select the initial centroid points for the K-means algorithm has advantages with the modified K-Means algorithm. with large data sets because it requires calculating the distance of every point with the first point of the given dataset as a very first step of the algorithm and sort it based on this distance.

The method is closely related to the approach of K-means clustering because it takes into account information reflecting the performance of the algorithm and systematic way to find initial centroid points which reduces the number of dataset scans and also will produce better accuracy in less number of iteration with the traditional algorithm. The method could be computationally expensive if used Author ming-chuan hung, jungpin wu+, jin-hua chang and don-lin yang [4] present an efficient algorithm to implement a k-means clustering that produces clusters comparable to slower methods. They partition the original dataset into blocks; each block unit, called a unit block, contains at least one pattern. They can locate the centroid of a unit block by using a simple calculation. All the computed CUBs form a reduced dataset that represents the original dataset. They have presented an efficient clustering algorithm based on the k-means method. They partitioned datasets into several blocks and used reduced versions of the datasets to compute final cluster centroids.

According to ahamed shafeeq b m 1 and hareesha k s 2 [3] they presents a modified K means algorithm with the intension of improving cluster quality. The K-means algorithm takes number of clusters (K) as input from user. The algorithm works well for the unknown data set with better results than K-means clustering. The k-means algorithm is well known for its simplicity and the modification is done in the method with retention of simplicity. The K-means algorithm takes number of clusters (K) as input from the use.

A. K. Malviya1, Vibhooti Singh2 [5] use the clustering technique of data mining in maintenance of software system using object oriented metrics. Evaluating the K-means clustering method by applying it to the commercial software system. The work of software maintenance for the sample data is being simulated on Matlab. The development of a methodology based on the K-means clustering data mining technique has been implemented on UIMS class data. The algorithm is able to decide the cluster with Good, Average and Bad conditions.

It proposes an algorithm which is of O(nk) time complexity. This time complexity is less than that of standard k-means algorithm. Experimental result shows that the running time is less than that of standard k-means. Thus it can conclude that the algorithm explained is feasible.

According to Cosmin Marian Poteras, Marian Cristian Mihˇaescu [7] an optimized version of the standard K-Means algorithm. The optimization refers to the running time and it comes from the observation that after a certain number of iterations, a small part of the data elements change their cluster, so there is no need to re-distribute all data elements. An optimized version of the K-Means algorithm. The optimization refers to the running time. It comes from the considerable reduction of the data space that is re-visited at each loop. The algorithm defines a ’border’ area made of those points that are close enough to the edge of their cluster so that the next centroids move could cause them to switch clusters.

Oyelade, O. J, Oladipupo, O. O, Obagbuwa, I. C [9] implemented k-mean clustering algorithm for analyzing students’ result data. The model was combined with the deterministic model to analyze the students’ results of a private Institution in%igeria which is a good benchmark to monitor the progression of academic performance of students in higher Institution for the purpose of making an effective decision by the academic planners. They provided a simple methodology to compare the predictive power of clustering algorithm and the Euclidean distance as a measure of similarity distance. Technique using k-means clustering algorithm and combined with the deterministic model in on a data set of private school results with nine courses offered for that semester for each student for total number of 79 students, and produces the numerical interpretation of the results for the performance evaluation.

According to ming-chuan hung, jungpin wu+, jin-hua chang and don-lin yang [12] they present an efficient algorithm to implement a k-means clustering that produces clusters comparable to slower methods. In our algorithm the original dataset into blocks; each block unit, called a unit block , contains at least one pattern. They can locate the CUB by using a simple calculation. All the computed CUBs form a reduced dataset that represents the original dataset. The reduced dataset is then used to compute the final centroid of the original dataset. They have presented an efficient clustering algorithm based on the k-means method. Datasets into several blocks and used reduced versions of the datasets to compute final cluster centroids.

According to Tapas Kanungo, Senior Member [16] they present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they call the filtering algorithm.Algorithm is easy to implement, requiring a kd-tree as the only major data structure.They establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases. Second, also present a number of empirical studies both on synthetically generated data and on real data sets from applications in color quantization, data compression, and image segmentation.They have presented an efficient implementation of Lloyd's k-means clustering algorithm, called the filtering algorithm.The algorithm is easy to implement and only requires that a kd-tree be built once for the given data points.

According to Unnati R. Raval1, Chaita Jani2 [17] they discusses the traditional K-means algorithm with advantages and disadvantages of it. It also includes researched on enhanced k-means proposed by various authors and it also includes the techniques to improve traditional K-means for better accuracy and efficiency.

There are two area of concern for improving K-means; 1) is to select initial centroids and 2) by assigning data points to nearest cluster by using equations for calculating mean and distance between two data points

The traditional K-means clustering is most used technique but it depends on selecting initial centroids and assigning of data points to nearest clusters. There are more advantages than disadvantages of the k-means clustering but it still need some improvements. It explains the techniques that improves the techniques for determining initial centroids and assigning data points to its nearest clusters with more accuracy with time complexity of O(n) which is faster than the traditional k-means.

According to adrian sergiu darabant1, laura darabant2 [2] An enhanced version for three clustering algorithms: hierarchical, k-means and fuzzy c-means applied in horizontal object oriented data fragmentation. Author is focusing in distributed object oriented database fragmentation and produce fragments for an OODB database based on the analysis of inter-class relationships and user queries (applications) running on the system. Each class extension is clustered and the quality of resulting fragments is then evaluated and compared. It represents three variations of some clustering algorithms with applications in database fragmentation.Beside the hierarchical clustering, the fuzzy c-means and k-means methods are scalable enough to be applied to large datasets.

According to Supreet Kaur, Dinesh Kumar [8] It necessitates the need to develop a real-time assessment technique that classifies these dynamically generated systems as being faulty/fault-free.These techniques include statistical method, machine learning methods, parametric models and mixed algorithms.The performance of the DBSCAN is evaluated for Java based Object Oriented Software system

.

from NASA Metrics Data Program data repository on the basis of fault proneness of the classes First, thirty nine metrics are used and later the worth of a subset of attributes is calculated and the number of metrics are reduced to eight. The metric values for the exemplars are used as Input and clusters are formed using DBSCAN, thereafter 10 fold cross validation performance of the system is recorded.

According to Deepali Virmani1,Shweta Taneja2,Geetika Malhotra3 [10] K-means is an effective clustering technique used to separate similar data into groups. N-K means clustering algorithm applies normalization prior to clustering on the available data calculates initial centroids based on weights. Experimental results prove the betterment of proposed N-K means clustering algorithm over existing K-means clustering algorithm in terms of complexity and overall performance. An efficient algorithm where we have first pre-processed our dataset based on normalization technique and then generated effective clusters.

According to Heba Ragab1, Amany Sarhan1, Al Sayed Sallam1, and Reda Ammar2 [11] Load balancing is a technique to distribute workload evenly across two or more computers, network links, CPUs, hard drives or other resources, in order to, get optimal resource utilization, maximize throughput, minimize response time and avoid overload. They introduce three clustering algorithms that obtain balanced clusters for homogeneous clustered with minimized communication cost. Load balancing is a computer networking methodology to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources.

István Gergely Czibula and Gabriela Şerban [13] they presenting an approach for improving systems design using clustering. Clustering is used in order to recondition the class structure of a software system. They have presented a approach, CARD, for improving systems design using clustering. They have introduced kRED algorithm, a k-means based clustering algorithm in order to obtain an improved structure of a software system. CARD proposes a list of refactoring that can be useful for assisting software engineers in their daily works of refactoring software systems.

According to Frantiek huka [14] the object oriented modelling in cluster analysis based on the Mjolner beta System and the BETA object oriented language. The model is designed in layers, which enables more flexibility and clear understanding. Object oriented modelling enables one to create a model that is more natural, and it is easy to extend or change its functionality. These features predetermine such a model for making experiments. Cluster analysis has many methods and techniques for helping to solve classification problems. Simplification of tasks is a principal way in which we are trying to cope with the complexity of models and programs. It is based on abstraction, i.e the removal or aggregation of details, components and/or relationships. To decrease mental effort, we can divide a system into substructures, which can be then treated as primitives at this level and defined as composites at the level below.

According to Tapas Kanungo, Senior Member [16] they present a simple and efficient implementation of Lloyd's k-means clustering algorithm, which they call the filtering algorithm. Algorithm is easy to implement, requiring a kd-tree as the only major data structure. They establish the practical efficiency of the filtering algorithm in two ways. First, we present a data-sensitive analysis of the algorithm's running time, which shows that the algorithm runs faster as the separation between clusters increases.

Second, also present a number of empirical studies both on synthetically generated data and on real data sets from applications in colour quantization, data compression, and image segmentation. They have presented an efficient implementation of Lloyd's k-means clustering algorithm, called the filtering algorithm. The algorithm is easy to implement and only requires that a kd-tree be built once for the given data points.

According to Amandeep Kaur Mann [18] there are different types of clusters: Well-separated clusters, Center-based clusters, Contiguous clusters, Density-based clusters, Shared Property or Conceptual Clusters. Predictive and the descriptive are the two main tasks of the data mining. Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms. Hierarchical clustering is the connectivity based clustering. Partitioning is the centroid based clustering, the value of k-mean is set. Density based clusters are defined as area of higher density then the remaining of the data set. Clustering is important in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups (clusters). Clustering can be done by the different no. of algorithms such as hierarchical, partitioning, grid and density based algorithms.

According to Dr. Sankar Rajagopal [19] he discuss on data mining process is to extracting valuable information from huge amounts of data. It is the process of discovering appealing knowledge from large amounts of data stored either in databases, data warehouses, or other information repositories. A necessary technique in data analysis and data mining applications is Clustering. A large amount of data is obtainable on the internet so it is complicated for the users to find out the pertinent data from this enormous data that is using a method clustering to solve these types of problems.

According to Periklis Andritsos [21] they present the state of the art in clustering techniques, mainly from the data mining point of view. They discuss the procedures clustering involves and try to investigate advantages and disadvantages of proposed solutions. The structure of this work is as follows: Section 2 outlines the stages commonly followed when performing clustering techniques. Section 3 discusses the different kinds of data we might have in hand, and metrics that define their similarities or dissimilarities. They described the process of clustering from the data mining point of view. We gave the properties of a "good" clustering technique and the methods used to find meaningful partitioning. At the same time, we concluded that research has emphasized numerical data sets, and the intricacies of working with large categorical databases are left to a small number of alternative techniques.

According to Tommi Karkkainen [22] Data clustering, by definition, is an exploratory and descriptive data analysis technique, which has gained a lot of attention, e.g., in statistics, data mining, pattern recognition etc. It is an explorative way to investigate multivariate data sets that contain possibly many different data types. These data sets differ from each other in size with respect to a number of objects and dimensions, or they contain different data types etc.They have given an introduction to the cluster analysis and reviewed partition-based clustering algorithms. A robust algorithm is introduced as an example. An important field of data mining applications is the data mining context. As the data mining concentrates on large real-world data sets, missing and erroneous values are often encountered.

Hierarchical Clustering

Hierarchical clustering involves creating clusters that l
all files and folders on the hard disk are organ
clustering, Divisive and Agglomerative.

A) Divisive method: In divisive or top-down clustering method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

B) Agglomerative method:
In agglomerative or bottom-up Clustering method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left.

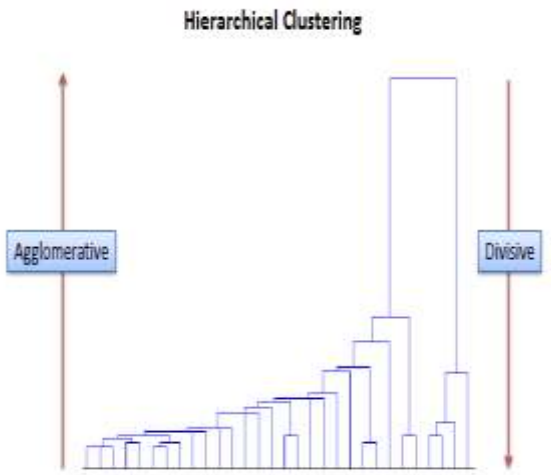
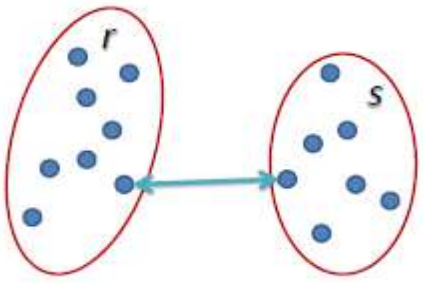


Figure 1

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between

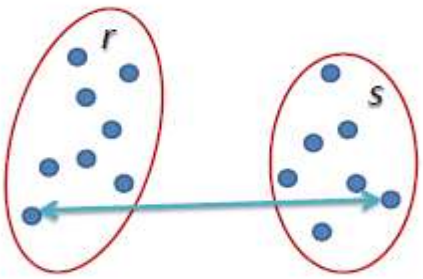
Single Linkage: In single linkage hierarchical clustering, the distance between two clusters is defined as the *shortest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two closest points.



$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Figure 2

Complete Linkage: In complete linkage hierarchical clustering, the distance between two clusters is defined as the *longest* distance between two points in each cluster. For example, the distance between clusters "r" and "s" to the left is equal to the length of the arrow between their two furthest points.

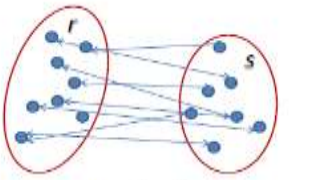


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Figure 3

Average Linkage:

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance b other cluster. For example, the distance between clusters "r length each arrow between connecting the points of one cluster to the other.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Figure 4

Example of Hierarical clustering

	A	B	C	D
A	0	0.71	5.66	3.61
B	0.71	0	4.95	2.92
C	5.66	4.95	0	2.24
D	3.61	2.92	2.24	0

There are 4 objects and we put each object in one cluster. The closest cluster is between cluster A and B with shortest distance of 0.71Thus,we group cluster A and cluster B into cluster(A,B).Distance between ungroup cluster will not change from the original distance matrix.

	AB	C	D
AB	0	?	?
C	?	0	?
D	?	?	0

	AB	C	D
AB	0	4.95	2.92
C	4.95	0,00	2.24
D	2.92	2.24	0

The closest cluster is between cluster C and D with shortest distance of 2.24 Thus, we group cluster C and cluster D into cluster(C,D)

Distance between cluster (A, B) and cluster (C,D) is

	AB	CD
AB	0	2.92
CD	2.92	0

4. Conclusion: In this paper we discuss two approaches of clustering. one is k-mean and second is Hierarical clustering. And also give the example of both clustering in object oriented concepts.

References

[1] Kedar b. sawant "Efficient Determination of Clusters in K-Mean Algorithm Using Neighbourhood Distance" International Journal of Emerging Engineering Research and Technology Volume 3, Issue 1, January 2015

[2] Adrian sergiu darabant1, laura darabant2 "Clustering methods in data fragmentation1" Romanian Journal of information Science and technology Volume 14, Number 1, 2011

[3] Ahamed shafeeq b m 1 and hareesha k s 2 "Dynamic Clustering of Data with Modified K-Means Algorithm" 2012 International Conference on Information and Computer Networks

[4] Ming-chuan hung, jungpin wu+, jin-hua chang and don-lin yang "An Efficient k-Means Clustering Algorithm Using Simple Partitioning*"journal of information science and engineering 21, 1157-1177 (2005)

[5] A.K. malviya1, vibhooti singh2 "The Role and Issues of Clustering Technique in Designing Maintainable Object Oriented System" A. K. Malviya et al. / International Journal on Computer Science and Engineering (IJCSE)
ISSN

[6] Sadhana Tiwari, Tanu Solanki " An Optimized Approach for k-means Clustering "9th International ICST Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine-2013)

[7] Cosmin Marian Poteras, Marian Cristian Mih˘aescu, "An Optimized Version of the K-Means Clustering Algorithm" Proceedings of the 2014 Federated Conference on Computer Science and Information Systems

[8] Supreet Kaur, Dinesh Kumar "Quality

[9] Oyelade, O. J, Ola dipupo, O. O, Obagbuwa, I. C "Application of k-Means Clustering algorithm for prediction of Students'sAcademicPerformance"(IJCSIS) International Journal of Computer Science and Information Security,
Vol. 7, _o. 1, 2010

[10] DeepaliVirmani1,ShwetaTaneja2,Geetika Malhotra3 " Normalization based K means Clustering Algorithm" Department of Computer Science, Bhagwan Parshuram Institute of Technology ,New Delhi

[11] Heba Ragab1, Amany Sarhan1, Al Sayed Sallam1, and Reda Ammar2 "Balanced Workload Clusters for Distributed Object Oriented Software "The International Arab Journal of Information Technology, Vol. 12, No. 4, July 2015.

[12] Ming-chuan hung, jungpin wu+, jin-hua chang and don-lin yang "An Efficient k-Means Clustering Algorithm Using Simple Partitioning" journal of information science

[13] István Gergely Czibula † and Gabri Şerban††, "Improving Systems Design Us a ClusteringApproach" IJCSNS Internatio Journal of Computer Science and Network Security, VOL.6 No.12, December 2006

[14] Franti□ek hu ka "object oriented approach in cluster analysis" Acta Electrotechnical Informatica No. 2, Vol. 3, 2003

[15] Kalpit G. Soni*1,2 and Dr. Atul Patel3 " Comparative Analysis of K-means and K-medoids Algorithm on IRIS Data" International Journal of Computational Intelligence Research ISSN 0973-1873 Volume 13, Number 5 (2017)

[16] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" ieee transactions on pattern analysis and machine intelligence, vol. 24, no. 7, july 2002