Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 7, July 2015, pg.453 – 458

REVIEW ARTICLE

Various Clustering Techniques in Software Engineering -A Review

Shalini Verma[1], Abhinav Mishra[2]

Student[1], Assistant Professor[2], Department of CSE

Chandigarh Engineering College, Landran, Punjab

Abstract: Software Engineering is a set of problem solving skills, instructions and methods applied upon a variety of domains to discover and create useful systems that is used to solve practical problems. There are many clustering techniques which are discussed in this paper.

Keywords: Clustering, software engineering, k-mean, clusters.

1. Introduction

Software is a not tangible device like computer programs and documentation. It is different from other tangible hardware device. Software Engineering is the discipline of computer science which follows engineering principles to create, operate, change and maintain of software components. Software Engineering is a set of problem solving skills, instructions and methods applied upon a variety of domains to discover and create useful systems that is used to solve practical problems. Software engineer is required to solve a problem or handle software engineering projects which evolve, create, build software and gives its behavior. Software engineers adopt approach regarding their work using some techniques, methodology and tools depending upon the resources available and problem to be solved. Software Engineering is the process of solving customer's problems by the systematic development and evolution of large, high quality software systems within cost, time and other constrains [11]. Software engineering is all about sequence of steps to produce the software, from its initial stage to its final stage. A software engineering is related to all the aspects that are used in the software production or create the software. Software is a generic term that is used for organizing the data and instructions that are collected to develop it.

The software is divided into the two categories: System Software and the Application Software. The system software is used to manage the hardware components, so that other software or user sees it as a functional unit. The software contains the operating system and some more utilities like disk formatting, file managers, display managers, etc. The application software used for accomplished the specific tasks. Application software may or may not contain the single program. Software is the program or set of programs. As in software many things are includes: as it consists of the programs, the complete documentation of that program, the procedure that is use to set up the software and the various operation of the software system. Software Engineering is a profession to provide high quality software products to its customers. It is an application of systematic, disciplined approach to development, operation, maintenance of software. Software consists of seven phases and these phases are called Software Development Life Cycle.

2. Review of Literature

In this paper [1] they explained about the reverse engineering concept is quite famous these days and related to recovery of software architecture. There are number of technique which as used in this paper to recover software architecture, one of them is clustering technique, which source the same component from software. Generally the component feature is vague. A group of same data element is known as clustering. This technique is as older and its used also in science and engineering. In simple words, identifying the number of data element, calculating similar coefficient and following the clustering method is called as clustering technique. The main function of the clustering technique for speedy and efficient recovery of software architecture by using fuzzy clustering technique. In this paper the major impact of this study shown that architecture recovery can be done better by fuzzy clustering instead of ordinary clustering. In this paper [2] they explained the adaptive fuzzy algorithm which is come along with the capability and adaptation. This adaptive caliber can be fulfill by using the tool of partition and consolidate it. The number of classes is the data set which requires the prior knowledge in fuzzy clustering algorithm. This new technique of algorithm can able to learn the number of classes continuously. Fuzzy mathematic provides the great accuracy results in clustering. The various techniques like k-mean, ISODATA, fuzzy C-mean and possibilistic C-mean algorithm is very effective where we require image segmentation. K-mean clustering identify the number of cluster continuously. The C-mean clustering and fuzzy C-mean clustering and new fuzzy clustering algorithm have an advantage when it combined with ISODATA. In this paper [3] they generate the idea about a technique which is based on image understanding and its analysis is called as remote sensing image segmentation. This paper is introduce the image analysis which required various technique i.e. Adaptive Genetic Algorithm (AGA) and alternative fuzzy C-Mean. The AGA identified the segmentation. The remote sensing images are always difficult because of they are equal grey pixel may be divide into different region of clustering. It is the better technique then the old technique which takes huge number of second. Whereas it take only needs few second. The segmentation process is the widely used technique in remote sensing images, which collect information, process of information and analysis. In this paper [4] they explained about Re-engineering software system is the recovery of software architecture and in software architecture recovery involves clustering. In this paper they guide us to introduce an approach that collectively clustering with matching technique to discover a decomposition which is well understood. Pattern matching is a technique under which architectural clues can be identified. All these clues are helpful to access an interclass similarity measure in clustering algorithm to produce the decomposition which is also known as final system decomposition. Adding a new updating in current existing software is always a challenging task but it also helpful to reduce the complexity in work. It is also necessary to keep update every error, patch or hack, for better performance of any software system or a software architecture. Architectural clue collect the source model is designed with proper information. In this paper [5] work represents ranking based method that improved K-means clustering algorithm performance and accuracy. In this they have also done analysis of K-means clustering algorithm, one is the existing K-means clustering approach

which is incorporated with some threshold value and second one is ranking method which is weighted page ranking applied on K-means algorithm, in weighted page rank algorithm mainly in links and out links are used and also compared the performance in terms of execution time of clustering. Proposed ranking based K-means algorithm produces better results than that of the existing k-means algorithm.

3. Clustering in Software Engineering*:*

Clustering technique defines classes and put objects which are related to them in one class on the other hand in classification objects are placed in predefined classes. Clustering means put the objects which have similar properties into one group and objects which have dissimilar properties into another [7]. In clustering, above threshold values objects can be placed in one cluster and values below into another cluster. Clustering has alienated the large data set into groups or clusters according to similarity in properties.

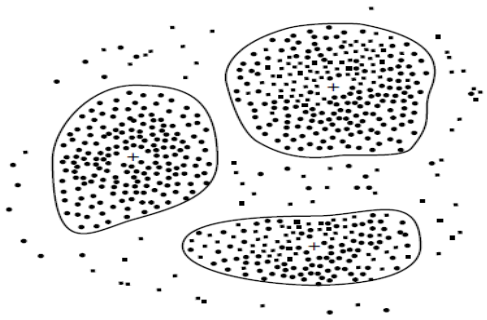

Fig 3: Cluster and Outlier

In above figure the dots which are outside the clusters represent outliers and there are clusters of object with similar properties.

3.2 Techniques of Clustering in Data Mining: Clustering is an unsupervised learning technique.

1. Partitioning clustering

2. Hierarchal clustering

3. Well shaped cluster

4. Density clustering

5. Centroid based cluster

6. K-mean clustering

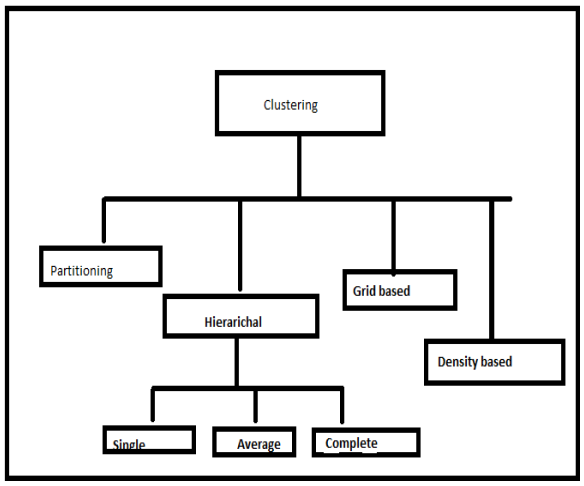

Fig..4: Techniques of Clustering

There are many techniques for clustering in data mining. These are as follow:

3.2.1 Partitioning Clustering: The general criterion for partitioning is a combination of high similarity of the samples inside of clusters with high dissimilarity between distinct clusters. Most partitioning methods are distance-based. These clustering methods are work well for finding spherical –shaped clusters in small to medium size databases [6].

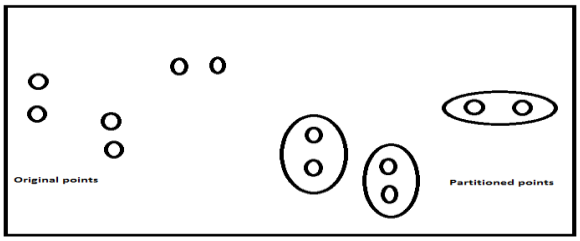

Fig.5: Partitioning Clustering

3.2.2 Density Based Clustering: Most partitioning methods cluster objects based on distance between objects. In these methods the cluster is continue to grow as long as the density in the neighbourhood exceeds some threshold [3].

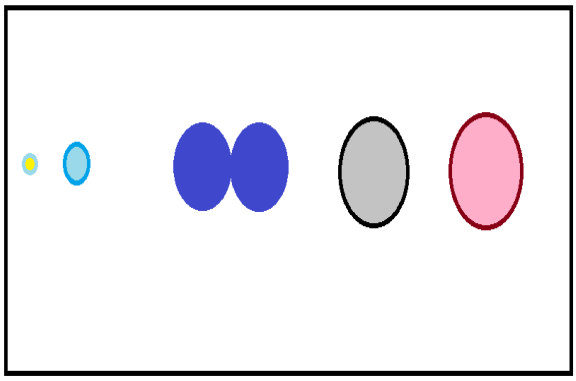

Fig.6: Density based Cluster

3.2.3 Grid Based Clustering: Grid based methods quantize the object space into a finite number of cells that form a grid structure. It is a fast method and is independent of the number of data objects and depends only on the number of cells in each dimension in the quantized space [8].

3.2.4 Hierarchical Methods*:* In this method hierarchical decomposition of the given set of data objects is created. It can be classified as being either agglomerative or divisive based on how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach starts with each object forming a separate group. Hierarchical algorithms create a hierarchical decomposition of the given data set of data objects. The hierarchical decomposition is represented by a tree structure, called dendrogram. It does not need clusters as inputs. In this type of clustering it is possible to view partitions at different level of granularities using different types of K. E.g. Flat Clustering [2].

It then merges groups close to one another until all the groups are merged into one. Divisive approach is top down approach starts with all the clusters in the same cluster and then in each iteration step a cluster is split into smaller clusters until each object is in one cluster.

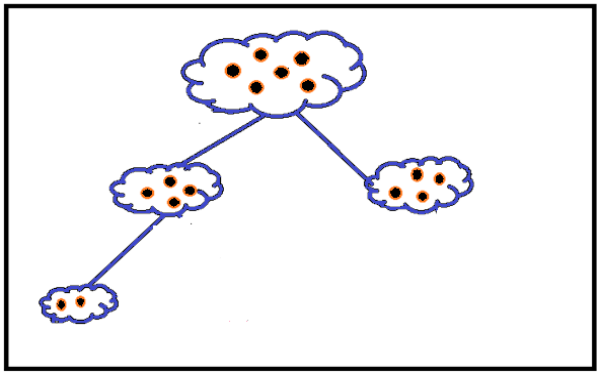

Fig.7: Hierarchal Clustering

3.2.5 Centre Based Cluster: A cluster is a set of objects. An object in cluster is more close to the central of a cluster which is similar not to the centre of any other cluster. A centroid which is an average of all points in cluster or a medoids which is most representative point in a cluster and often the centre of a cluster.

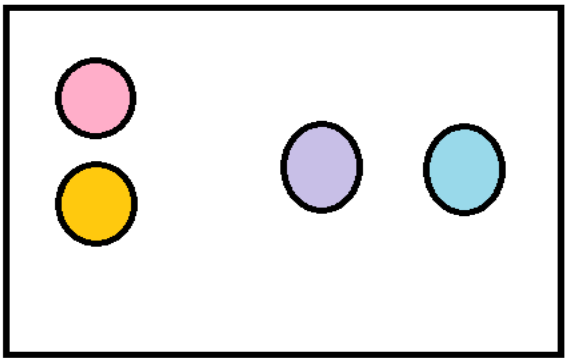

Fig.8:Centre based clustering

3.2.6 Well shaped Cluster: A cluster is a package of nodes in which any node in a cluster is closer or more similar to every other node in the same cluster than to any node not in the cluster. Sometimes threshold can be used to specify similarity or closeness between the nodes in cluster [9].

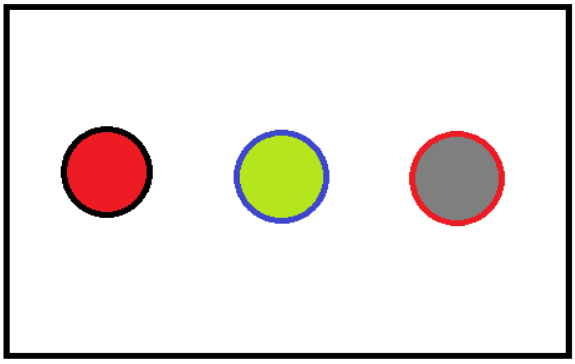

Fig.9: Well shaped Cluster

Instead of these techniques there are more types of cluster also there like conceptual clusters, well separated cluster etc.

3.2.7 K-Mean Clustering: The k-means clustering algorithm is the basic algorithm which is based on partitioning method which is used for many clustering tasks especially with low dimension datasets. It uses k as a parameter, divide n objects into k clusters so that the objects in the same cluster are similar to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, (*C1 …… Ck*), such that the sum of the squared distances of each data point, *xi* ,*1* $\leq i \leq n$, *to* its nearest cluster centre *Cj*, $1 \leq j \leq k$, is minimized. First, the algorithm randomly selects the k objects, each of which initially represents a cluster mean or centre. Then, each object *xi* in the data set is assigned to the nearest cluster centre i.e. to the most similar centre [4]. The algorithm then computes the new mean for each cluster and reassigns each object to the nearest new centre. This process iterates until no changes occur to the assignment of objects.

4. Conclusion

In this paper, it is concluded that clustering is technique in which large datasets are divide in to small datasets in this way that objects and items with having similar properties into one group and objects having dissimilar properties into another. There are number of algorithms that work well. In this paper we have reviewed various types of clustering which are well suited in different types of environment. By using these clustering techniques we can improve accuracy of the architecture.

References

[1] Lingming Zhang, Ji Zhou, Dan Hao ,Lu Zhang, Hong Mei" *Prioritizing JUnit Test Cases in Absence of Coverage Information*" IEEE 2009.

[2] Paolo Tonella, Paolo Avesani, Angelo Susi" *Using the Case-Based Ranking Methodology for Test Case Prioritization*". 22nd IEEE International Conference on Software Maintenance (ICSM'06),2009.

[3] Zheng Li, Mark Harman, and Robert M. Hierons" *Search Algorithms for Regression Test Case Prioritization*" IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 33, NO. 4, APRIL 2007.

[4] Amar Singh and Navot Kaur, "To Improve the Convergence Rate of K-Means Clustering Over K-Means with Weighted Page Rank Algorithm," *International journal of Advanced Research in Computer Science and Software Engineering*, Volume 3, Issue 8, August 2012.

[5] K. A. Abdul Nazeer, M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering , Vol IWCE 2009, July 1 - 3, 2009, London, U.K

[6] Batagelj,V., Mrvar,A.,andZaversnik,M., ",*Partitioning approaches toclusteringin graphs,* Pr Drawing'1999, LNCS, 2000, pp. 90-97.

[7] Ertoz, L., Steinbach, M., and Kumar, V., *"Finding clusters of different sizes, shapes, and densitie dimensional data"*, In Proc. of SIAM DM'03.

[8] Ester, M., Kriegel, H.P., Sander,J., and Xu, X., *" A density-based algorithm for discovering clusters databases with noise",* in Proc. of 2nd Int. Conf. on Knowledge Discovery and Data Mining(KDD-96),AAAI Press, 1996, pp. 226-231.

[9] Fayyad, U. and Grinstein,G., *"Information Visualization in Data Mining and Knowledge Discovery*", M 2001, pp. 182-190.

[10] Han, J., Kamber, M., and Tung, A. K. H., "Spatial clustering methods in (eds.), *Geographic Data Mining and Knowledge Discovery*, TaylorandFrancis, 2001.

[11] Harel, D.andKoren, Y., *"Clustering spatial data using random walks",* In Proc. 7th and Data Mining(KDD-2001),ACM Press, New York, pp. 281-286

[12] Satoshi Takumi and Sadaaki Miyamoto, "*Top-down vs Bottom-up methods of Linkage for Asymmetric Agglomerative Hierarchical Clustering*",International Conference on granular Computing, 2012

