# Clustering Techniques in Data Mining For Improving Software Architecture: A Review

Parneet Kaur
GNDU University, Amritsar (143005),
PUNJAB

Kamaljit Kaur
GNDU University, Amritsar (143005),
PUNJAB

## ABSTRACT

Data mining is a set of problem solving skills, instructions and methods applied upon variety of domains to discover and create useful systems that are used to solve practical problems. 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. There are many clustering techniques for the improvement of architecture which are discussed in this paper. This paper also gives comparative study of clustering techniques and addresses benefits and limitations of clustering techniques.

## General Terms

Data Mining, Software Re-engineering, Clustering.

## Keywords

Clustering, Software Engineering, k-means, Outliers.

## 1. INTRODUCTION

Software is not a tangible device like computer programs and documentation. It is different from other tangible hardware devices. Data mining is the discipline of computer science which follows engineering principles for creating, operating, changing and maintaining software components. Data mining is a set of problem solving skills, methods and instructions which are applied upon a variety of domains to discover and create useful systems that are used to solve practical problems. A Software engineer requires solving a problem or handling Data mining projects which evolve, create, build software and gives its behaviour. Software engineers adopt approaches regarding their work using some techniques, methodology and tools depending upon the resources available and problem to be solved. Data mining is the process of solving customer's problems by developing large, high quality software systems within cost, time and other constrains [11]. Data Mining is all about sequence of steps to produce the software, from its initial stage to its final stage. It is related to all the aspects that are used in the software production or creation of 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 and display managers. Application software may or may not contain the single program. Software is the program or set of programs. Software includes many things such as it consists of the programs, the complete documentation of that program, the procedure that is used to set up the software and the various operation of the software system. Data mining is a profession to provide high quality software products to its customers. It is an application of systematic, disciplined approach for development, operation, maintenance of software. Software consists of seven phases and these phases are called Software Development Life Cycle [10].

## 2. REVIEW OF LITERATURE

In this paper [1] it is explained that the reverse engineering concept is quite popular and it is related to recovery of software architecture. There are number of techniques which are used in this paper to recover software architecture and clustering is one of them which source the same component from software. The component feature is generally vague. A group of same data elements is known as cluster. This technique is older and is used in science and engineering. In simple words, identifying the number of data elements, calculating similar coefficient and following the clustering method is known as clustering technique. Fuzzy clustering technique is used to achieve the main function of clustering technique which is used for efficient and speedy recovery of software architecture. In this paper the major impact of study shows that architecture recovery can be done in a better way by fuzzy clustering instead of ordinary clustering. The adaptive fuzzy algorithm is explained in this paper [2], which comes along with the adaptation and capability. This adaptive caliber can be fulfilled by using the tool of partition. Prior knowledge is required by the number of classes in the data set in case of fuzzy clustering algorithm. This new technique of algorithm is able to learn the number of classes continuously. Great accuracy results are provided by Fuzzy mathematic when it is used in clustering. Many techniques like k-means, ISODATA, fuzzy C-mean and possibilistic C-mean algorithm are very effective where we require image segmentation. The number of clusters is identified continuously by k-means approach. The fuzzy C-mean, C-mean clustering and new fuzzy clustering algorithm have an advantage when it is combined with ISODATA. In this paper [3] a technique based on image understanding and its analysis called as remote sensing image segmentation is presented. This paper introduces the image analysis which requires various techniques like Adaptive Genetic Algorithm (AGA) and alternative fuzzy C-Mean. The AGA identifies the segmentation. It is difficult to segment remote sensing images because they have equal grey pixels and they may be divided into different regions of clustering. It is the better technique then the old technique which requires large computational time whereas, it take only few seconds. The segmentation technique is widely used in remote sensing images, which collects information, processes information and analysis it. This paper [4] explains about Re-engineering software system which deals with the recovery of software architecture which involves clustering. In this paper the author also guides us to introduce an approach that collective clustering with matching technique discovers a decomposition which is well understood. Architectural clues can be identified by using pattern matching technique. All these clues are helpful for accessing an interclass similarity measure in clustering algorithm for producing the decomposition called as final system decomposition. It is a challenging task to add new updating in current software but it also helpful to reduce the complexity in work. It is necessary to keep update patch or hack and every error for better performance of any software system or software architecture. Architectural clue to collect the source model is designed with proper information. In this paper [5] ranking based method is presented, that improves performance and accuracy of k-means clustering algorithm. In this paper, k-means clustering technique is analyzed, 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 there performance is compared in terms of execution time. Better results are given by proposed ranking based k-means algorithm than that of the existing k-means algorithm. This paper [23] presents two clustering techniques, Modex and seed –Detective which produce high quality clusters by using good initial seeds in k-Means. ModEx addresses limitations of Ex-Detective technique and is the modified version of Ex-Detective. Simple K-Means and ModEx are combined to produce Seed Detective technique. Seed-Detective uses ModEx for the production of good quality initial seeds and give these seeds as input to k-means which leads to the production of high quality clusters. This paper [19] gives distributive agglomerative cluster based anomaly detection algorithm (DACAD), which is based on kNN approach. This approach improves the reliability in wireless sensor networks by detecting the faulty readings in wireless sensor networks. In this paper [15] partitioning method named as Clustering Large Applications (CLARA) is used for identification of métiers of the Northern Spanish coastal bottom pair trawl fleet. The partitioning method CLARA is chosen for clustering large datasets because this partitioning technique is specifically designed for managing very large data sets. This paper [20] gives medical image segmentation technique by using k-means clustering integrated with Fuzzy C-means algorithm. The proposed technique gets benefits of the k-means clustering for image segmentation in the aspects of minimal computation time and it get advantages of the Fuzzy C-means in the aspects of accuracy. In this paper [16] an improved semi-supervised clustering algorithm which is based on SCMD algorithm is presented. This proposed algorithm easily deals with multi-density problems which includes both inter and intra cluster density and yields superior performance with fewer constraints. In this paper [22] density based spatial clustering (DBSCAN) is used for analyzing Hotspot distribution in case of forest and land fires by determining the areas with high density. This algorithm can easily find arbitrary shaped clusters and efficiently handles noise. This paper [17] gives distributed clustering algorithm based on heterogeneous cloud computing known as HiClus. HiClus efficiently builds an adaptive distributed tree in the cloud for utilizing computational resources of both general-purpose computing on graphics processing units and computational processing unit. This technique has less computational time and achieves better load balancing as compared to Map Reduce algorithms. In this paper [21] Self Organized Maps (SOM) are used for classification of consumers according to their water consumption. Such analysis can be promising for the automatic classification of water consumers, based on urban water demand data. The preliminary analysis of urban water consumption data is conducted for the classification of users into different categories on the basis of their water consumption. This technique automatically determines the number of clusters and can be used even when the dataset is not real. This paper [18] illustrates weakness of k-medoids algorithms, in regard to the optimality and feasibility of the solutions. So, two variants are exposed for partitioning around medoids for data with balance restrictions over the number of objects present in each group keeping the feasibility and optimality of the solution. In the first algorithm, the ideas of k-medoids are considered and extend it with a recursive constructive function to find balanced solutions. The second algorithm searches for solutions taking into account a balance between compactness and the cardinality of the groups (multi objective).

## 3. CLUSTERING IN DATA MINING

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 putting the objects which have similar properties into one group and objects having dissimilar properties into another group [7]. Threshold value is defined and values of objects above threshold are 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. Outliers are the data points which are present outside the clusters. In figure 1, the dots which are outside the clusters represent outliers and there are cluster of objects with similar properties.

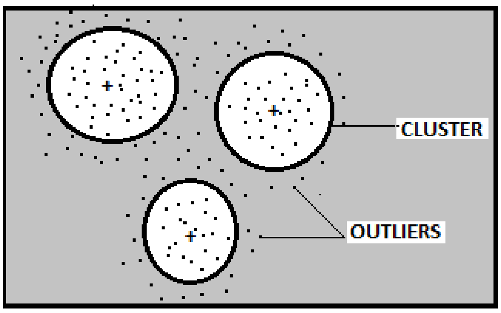

**Fig 1: Clusters and Outliers**

### 3.1 Partitioning Clustering

Partitioning clustering is based on the general criterion of combining high similarity of the samples inside of clusters with high degree of dissimilarity among 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].

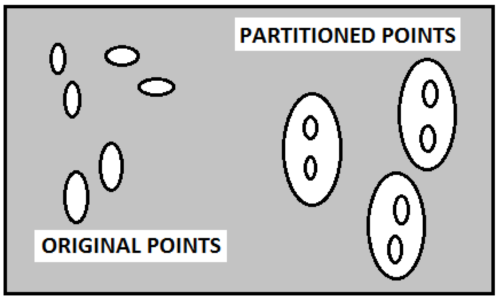

**Fig 2: Partitioning Clustering**

## 3.2 Density Based Clustering

Most partitioning methods cluster objects based on distance between objects. In these methods the cluster continues to grow as long as the density in the neighbourhood exceeds some threshold [3].

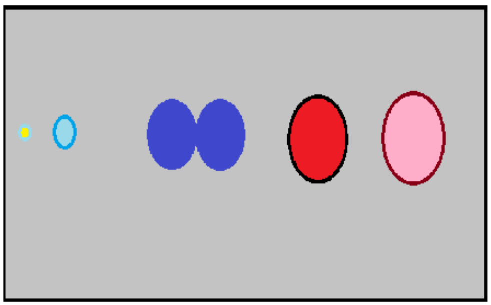

**Fig 3: Density Based Clusters**

## 3.3 Grid Based Clustering

In Grid based methods, the object space is quantized 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 is dependent on only the number of cells present in each dimension in the quantized space [8].

## 3.4 Hierarchical Methods

In this method hierarchical decomposition of the given set of data objects is created. It can be classified into two categories named as agglomerative or divisive, on the basis that how hierarchical decomposition is formed. Agglomerative approach is the bottom up approach starting with each object forming a separate group. Hierarchical algorithms create a hierarchical decomposition of the data set containing data objects. It 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 [12]. Then the groups close to one another are merged, until all the groups are merged into one.

Divisive approach is top down approach which 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 are in one cluster.

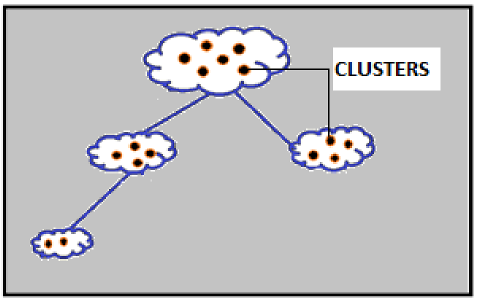

**Fig 4: Hierarchical Clustering**

## 3.5 Centre Based Clustering

A cluster is a set of objects. An object in cluster is more close to the centre of a cluster which is not similar to the centre of any other cluster. A centroid is an average of all points in cluster or a medoids. It is the most representative point in a cluster and often the centre of a cluster [14].

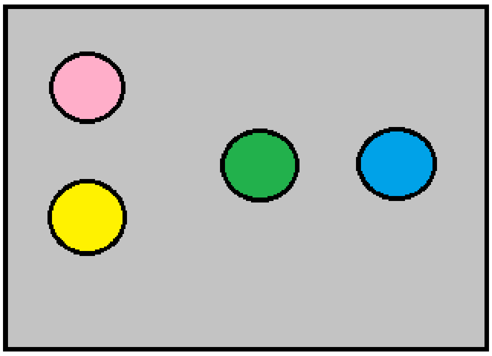

**Fig 5: Centre Based Clustering**

## 3.6 Well Shaped Clusters

A cluster is a package of nodes in which any node in a cluster is more similar or closer to every other node of the cluster in which it is present than to any node not in the cluster. Sometimes threshold can be used to specify closeness or similarity among the nodes in cluster [9].

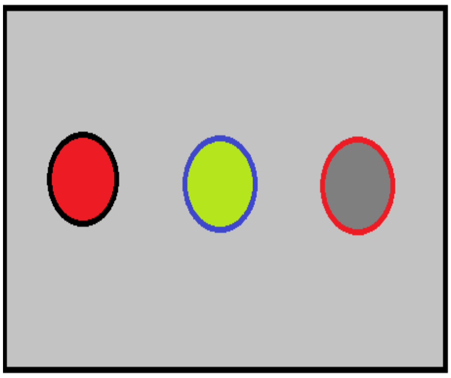

**Fig 6: Well Shaped Clusters**

## 3.7 K- Means 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 within the cluster are identical to each other but dissimilar to other objects in other clusters. The algorithm attempts to find the cluster centres, ($C1 \ldots\ldots Ck$), 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]. Then new mean is computed for each cluster and each object is reassigned to the nearest new centre. This process iterates until no changes occur to the assignment of objects [13].

## 4. SUMMARY OF CLUSTERING APPROACHES

This section summarizes clustering techniques reviewed earlier by giving the benefits and limitations of clustering techniques. The limitation of one approach may get overcomed by another approach.

**Table 1. Clustering Techniques**

| Author (Year) | Technique | Benefits | Limitations |
|---|---|---|---|
| **Md Anisur Rahman**<br>**(2015)** | Modex, Seed detective & k Means clustering | High quality seeds are produced for input to k means clustering. | Clusters members are to be defined and there is occurrence of noise. |
| **N.Chitra Devi**<br>**(2011)** | Distributive agglomerative cluster based anomaly detection algorithm<br>(DACAD) | Support applications with heterogeneous data. | Normal data may be detailed as outlier. |
| **B. Bernábe-Loranca**<br>**(2014)** | K medoids algorithm and Partitioning around medoids<br>(PAM) | The partitioning around medoid gives satisfactory results when the instances are small. | Present weakness in regard to the optimality and feasibility of the solutions in the presence of additional restrictions. |
| **Eman Abdel – Maksoud**<br>**(2015)** | k-Means clustering integrated with Fuzzy C Means algorithm<br>(KIFCM) | Minimal computational time, fast technique and and has advantages of Fuzzy C Means in the aspects of accuracy. | Does not give 3D evaluation of brain tumor and is less efficient. |
| **Xiaoyun Chen**<br>**(2012)** | Improved Semi supervised clustering algorithm<br>(SCMD) | Deals with multi density problems and yields superior performance with fewer constraints. | Cannot detect the cluster in the clustering process having no constraints. |
| **Muhammad Usman**<br>**(2015)** | Density based spatial clustering<br>(DBSCAN) | Find arbitrary shaped clusters and easily handles noise. | Not suitable for environment with different densities. |
| **Chun-Chieh Chen**<br>**(2015)** | Highly scalable density based clustering with heterogeneous cloud<br>(HiClus) | Scale up better, Use less clustering time and achieve better load balancing than existing map reduce algorithms. | Hiding memory access latencies becomes very important in HiClus. |
| **Chrysi laspidou**<br>**(2015)** | Self organized maps<br>(SOM) | Works even when the data is not real time and frequent and automatically determines the number of clusters. | Efficiency may degrade in case of large datasets. |
| **Jose Castro**<br>**(2010)** | Clustering large applications<br>(CLARA) | Easily manage very large datasets. | Overlapping of clusters may occur. |

## 5. CONCLUSION

In this paper, it is concluded that clustering is technique in which large datasets are divided into small groups. The objects and items having similar properties are grouped into one group and objects having dissimilar properties into another. There are number of algorithms that work well and by using clustering technique, the architecture of the system can be improved. In this paper review of clustering techniques is done and there benefits and limitations are addressed. The limitations and issues arising in clustering algorithms may be beneficial for future researchers.

## 6. 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 Navjot 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, and Zaversnik.M, "Partitioning approaches to clustering in graphs," Pr Drawing'1999, LNCS, 2000, pp. 90-97.

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

[8] Ester, M., Krieger, 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, Taylor and Francis, 2001.

[11] Harel, D. and Koren, 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.

[13] Kiran Agrawal, Ashish Mishra, "Improved K-MEAN Clustering Approach for Web Usage Mining", ICETET, 2009, Emerging Trends in Engineering & Technology, International Conference on, Emerging Trends in Engineering & Technology, International Conference on 2009, pp. 298-300.

[14] Rudolf Scitovski, Tomislav Marošević, "Multiple circle detection based on center-based clustering, Pattern Recognition Letters," Volume 52, 15 January 2015, Pages 9-16, ISSN 0167-8655.

[15] José Castro, Antonio Punzón, Graham J. Pierce, Manuel Marín, Esther Abad, Identification of métiers of the Northern Spanish coastal bottom pair trawl fleet by using the partitioning method CLARA, Fisheries Research, Volume 102, Issues 1–2, February 2010, Pages 184-190, ISSN 0165-7836.

[16] Xiaoyun Chen, Sha Liu, Tao Chen, Zhengquan Zhang, Hairong Zhang, An Improved Semi-Supervised Clustering Algorithm for Multi-Density Datasets with Fewer Constraints, Procedia Engineering, Volume 29, 2012, Pages 4325-4329, ISSN 1877-7058.

[17] Chun-Chieh Chen, Ming-Syan Chen, HiClus: Highly Scalable Density-based Clustering with Heterogeneous Cloud, Procedia Computer Science, Volume 53, 2015, Pages 149-157, ISSN 1877-0509.

[18] B. Bernábe-Loranca, R. Gonzalez-Velázquez, E. Olivares-Benítez, J. Ruiz-Vanoye, J. Martínez-Flores, Extensions to K-Medoids with Balance Restrictions over the Cardinality of the Partitions, Journal of Applied Research and Technology, Volume 12, Issue 3, June 2014, Pages 396-408, ISSN 1665-6423.

[19] N. ChitraDevi, V. Palanisamy, K. Baskaran, S. Prabeela, A Novel Distance for Clustering to Support Mixed Data Attributes and Promote Data Reliability and Network Lifetime in Large Scale Wireless Sensor Networks, Procedia Engineering, Volume 30, 2012, Pages 669-677, ISSN 1877-7058.

[20] Eman Abdel-Maksoud, Mohammed Elmogy, Rashid Al-Awadi, Brain tumor segmentation based on a hybrid clustering technique, Egyptian Informatics Journal, Volume 16, Issue 1, March 2015, Pages 71-81, ISSN 1110-8665.

[21] Chrysi Laspidou, Elpiniki Papageorgiou, Konstantinos Kokkinos, Sambit Sahu, Arpit Gupta, Leandros Tassiulas, Exploring Patterns in Water Consumption by Clustering, Procedia Engineering, Volume 119, 2015, Pages 1439-1446, ISSN 1877-7058.

[22] Muhammad Usman, Imas Sukaesih Sitanggang, Lailan Syaufina, Hotspot Distribution Analyses Based on Peat Characteristics Using Density-based Spatial Clustering, Procedia Environmental Sciences, Volume 24, 2015, Pages 132-140, ISSN 1878-0296.

[23] Md Anisur Rahman, Md Zahidul Islam, Terry Bossomaier, ModEx and Seed-Detective: Two novel techniques for high quality clustering by using good initial seeds in K-Means, Journal of King Saud University - Computer and Information Sciences, Volume 27, Issue 2, April 2015, Pages 113-128, ISSN 1319-1578.