A REVIEW ON DATA CLUSTERING AND K-MEANS CLUSTERING ALGORITHM FOR FAULT DETECTION

Dr.Satwinder Singh,[1] Sukhjeet kaur[2]
[1]Assistant Professor, [2]Research Scholer
[1,2]Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib, (Punjab), India
[1]satwinder.singh@bbsbec.ac.in, [2]sukhjeet025@gmail.com

***ABSTRACT* : Clustering is the task of grouping & organizing a set of similar objects together. The set of organized objects with the help of clustering is known as clusters. Clustering together with software artifacts provides an automatic technique for discovering high level abstract entities within a system. K-means clustering algorithm which came into existence in 1955 is most efficient, understandable and easy to grouped similar behavioral objects. After that, many algorithms have evolved over the years but k-means is still being widely used. The reason is complications in designing of a clustering algorithm and ill-posed problem of clustering. The history and brief over view of popular clustering methods, key issues in designing clustering algorithms and the useful researches over the clustering will be discussed in this work. The implementation and analysis of k-means clustering for fault prediction in software modules can be better done using WEKA. The open source Object-Oriented metrics are available for defect prediction.**

***KEYWORDS* : Clustering, K-means, Developments, Weka, Object-Oriented Metrics, fault prediction.**

I. INTRODUCTION

The growth of data increases day by day, so the different data is available. The images and videos contain large amount of data. It is estimated by Gantz[1] that the digital universe consumed approximately 281 exabytes in 2007, and it is projected to be 10 times that size by 2011 (1 exabyte is 101,000,000 terabytes).Most of this data is unstructured, which lead to the difficulty in analyzing it. Tukey [2] broadly classified data analysis techniques into two major types: (i) exploratory or descriptive, meaning the researcher is interested in understanding general characteristics or structure of the high-dimensional data but pre-specified models or hypotheses are not required, and (ii) confirmatory or inferential meaning that the researcher is interested in confirming the validity of a hypothesis/model or a set of assumptions given the available data. The data clustering is a very good technique for organizing and analyzing the data.

A. Data clustering

A differentiation is made between learning problems by Duda [3] that are (i) classification (supervised) or (ii) clustering (unsupervised).Clustering is a more complicated problem than classification. In general, in classification there has a set of predefined classes and want to know which class a new object belongs to. Clustering tries to group a set of objects and find whether there is some relationship between the objects.

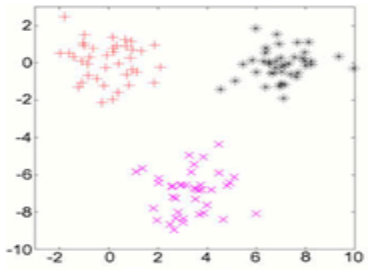

Fig.1 Supervised

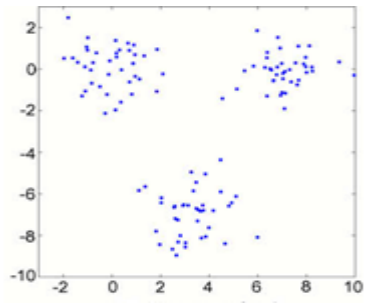

Fig.2 Unsupervised

The aim of data clustering, also known as cluster analysis, is to discover the natural grouping(s) of a set of patterns, points, or objects. Webster (Merriam-Webster Online Dictionary, 2008) defines cluster analysis as ‘‘a statistical classification technique for discovering whether the individuals of a population fall into different groups by making quantitative comparisons of multiple characteristics.”

1) Importance of data clustering:

Cluster analysis is applicable in any discipline that contains analysis of multivariate data. A search via Google Scholar (October 2014) found 2,600,000 entries with the words data and clustering that appeared in 2014.It is complex to comprehensively record the innumerable scientific fields and applications that have used clustering techniques and many of published algorithms. Image segmentation, a significant dilemma in computer vision, can be coined as a clustering problem. Data clustering is applicable for the following main intents: Underlying structure, Natural classification and Compression.

2) Background development:

According to JSTPR [4], data clustering first appeared in the title of a 1954 article dealing with anthropological data. Data clustering is acknowledged as Q-analysis, typology, clumping, and taxonomy as well. Major approaches to clustering will be briefly reviewed .Frank and Todeschini [5] defined the Jarvis–Patrick algorithm that states the similarity between a pair of

points as the number of common neighbors they share. Dempster [6] proposed the EM algorithm, is usually applied to infer the parameters in mF4ixture models. Agrawal [7] defined CLIQUE to find subspaces in the data with high-density clusters. The representation of the data points as nodes in a weighted graph is done using Graph theoretic clustering, also known as spectral clustering.
Clustering algorithms can be subdivided as hierarchical and partitional. K-means is the simplest and most popular partitional algorithm.

II. LITERATURE REVIEW

B. k-means clustering:
The main steps of K means algorithm as defined by Jain and Dubes [8] are: 1.Select an initial partition with K clusters; repeat steps 2 and 3 until cluster membership stabilizes.2. Generate a new partition by assigning each pattern to its closest cluster center.3.Compute new cluster centers. The K-means algorithm depends upon three user-specified parameters: number of clusters K, cluster initialization, and distance metric. Choosing value of K is most censorious. There are different extensions of K-means algorithm. The enlarged heuristics are tackled by some of these extensions that involve the minimum cluster size and merging and splitting clusters. The two popular alternatives of K-means in pattern recognition literature are ISODATA proposed by Ball and Hall [9] and FORGY Forgy [10]. Fuzzy c-means is an extension of K-means where each data point can be a member of multiple clusters with a membership value (soft assignment), which was proposed by Dunn [11] and later improved by Bezdek [12].

C. Difficulties and challenges for users regarding clustering:
The elementary challenges associated with clustering were highlighted in Jain and Dubes [13] are as follows:
(a) What is a cluster?
(b) What features should be used?
(c) Should the data be normalized?
(d) Does the data contain any outliers?
(e) How do we define the pair-wise similarity?
(f) How many clusters are present in the data?
(g) Which clustering method should be used?
(h) Does the data have any clustering tendency?
(h) Does the data have any clustering tendency?
Some of these challenges will be highlighted and illustrated. The making of clusters becomes easy while the data representation is good and the clusters become compact and isolated. Then even a simple clustering algorithm can easily find those clusters. The purpose of grouping plays an important role in the data representation. Most methods for automatically determining the number of clusters cast it into the problem of model selection. Mostly, the different values of k are used by the clustering algorithms. Based on the predetermined criterion, the best value of k is selected. The minimum message length abbreviated as MML, criteria proposed by Wallace and Boul ton; Wallace and Freeman[14] in conjunction with the Gaussian mixture model (GMM) to estimate K. Cluster validity indices can be defined based on three different criteria: internal, relative, and external, was used by Figueiredo and Jain[15].

D. Comparison of clustering algorithms:
Even if the different clustering algorithms applied on the same data, often the results can be entirely different partitions. FORGY, ISODATA, CLUSTER, and WISH are partitional algorithms that minimize the squared error criterion.MST (minimum spanning tree) works as a single-link hierarchical algorithm, and JP is a nearest neighbor clustering algorithm. A partition can be generated by specifying a threshold on the similarity by applying a hierarchical algorithm. Clustering algorithms were formally analyzed by Fisher and vanNess [16] with the objective of comparing them and providing guidance in choosing a clustering procedure. A set of admissibility criteria for clustering algorithms was defined by them that test the sensitivity of clustering algorithms with respect to the changes that do not alter the essential structure of the data. A clustering is called A- admissible if it satisfies criterion A.

E. Trends in clustering
A diverse set of data is created by Information explosion along with the large amounts of data. The data created is both structured and unstructured. The structure in the objects is ignored by most of the clustering approaches and a feature vector based representation for both structured and unstructured data is used by them. The traditional view of data partitioning based on vector- based feature representation does not always serve as an adequate framework. A brief summary of some of the recent trends in data clustering is presented below.

1) Developments:
Fred and Jain [17] stated that the development of ensemble methods for unsupervised learning has been motivated by the success of ensemble methods for supervised learning. By taking different looks at the same data different partitions of the same data can be generated, which is the principle idea. It is possible to obtain a good data partitioning while the clusters are not compact and well separated by combining the resulting partitions. Then a co-occurrence matrix that provided a good separation of the clusters was used to combine these partitions.
The data based on the new pair-wise similarity is used to obtain the resulting clustering. Many different ways of generating a clustering ensemble and then combining them are available. For example, multiple data partitions can be generated by: (i) applying different clustering algorithms, (ii) applying the same clustering algorithm with different values of parameters or initializations, and (iii) combining of different data representations (feature spaces) and clustering algorithms.
Semi-supervised clustering:
Chapelle [18] stated that clustering algorithms are said to be operating in a semi-supervised mode that utilize such side information. There are two open questions: (i) how should the side information be specified? And (ii) how is it obtained in practice? A must-link constraint specifies that the point pair connected by the constraint belong to the same cluster. On the other hand, a cannot-link constraint specifies that the point pair connected by the constraint do not belong to the same cluster. Other approaches for including side information include (i) ‘‘seeding”, where some labeled data is used along with large

amount of unlabeled data for better clustering stated by Basu[19] and (ii) methods that allow encouraging or discouraging some links stated by Law[20] and Figueiredo[21].

Large-scale clustering:
The challenge of clustering millions of data points that are represented in thousands of features is addressed by large-scale data clustering. The application of large-scale data clustering to content-based image retrieval is reviewed below.
The goal of Content Based Image Retrieval (CBIR) is to retrieve visually similar images to a given query image. There is not much success in this topic instead of the study for the past 15 years. Datta [22] stated that a 2008 survey on CBIR highlights the different approaches used for CBIR through time. Recent approaches for CBIR use key point based features. On the other hand, text retrieval applications are much faster. It takes about one-tenth of a second to search 10 billion documents in Google. A large number of clustering algorithms have been developed to efficiently handle large-size data sets. Most of these studies can be classified into four categories:
1. Efficient Nearest Neighbor (NN) Search, 2.Data summarization, 3.Distributed computing, 4.Incremental clustering and 5.Sampling based clustering

Multi-way clustering:
A combination of related heterogeneous components forms objects or entities that have to be clustered. The objects can be converted into a pooled feature vector of its components but it is not a natural representation of the objects and it may result in poor clustering performance. Hartigan[23] and Mirkin[24] defined Co-clustering as it aims to cluster aims to cluster both features and instances of data simultaneously to identify the subset of features where the resulting clusters are meaningful according to certain evaluation criterion. Bi-dimensional clustering double clustering, coupled clustering, or bimodal clustering are its other names. Bekkerman [25] stated that the co-clustering framework was extended to multi-way clustering into cluster a set of objects by simultaneously clustering their heterogeneous components. The problem is much more challenging because of different relationships. *Heterogeneous data:* Heterogeneous data refers to the data where the objects may not be naturally represented using a fixed length feature vector. Rank data: Consider a dataset generated by ranking of a set of n movies by different people; only some of the n objects are ranked movies by different people; only some of the n objects are ranked. The task is to cluster the users whose rankings are similar and also to identify the 'representative rankings' of each group. Dynamic data: Dynamic data, as opposed to static data, can change over the course of time e.g., blogs, Web pages, etc. As change over the course of time e.g., blogs, Web pages, etc. A data stream is a kind of dynamic data that is transient in nature, and cannot be stored on a disk. Graph data: Several objects, such as chemical compounds, protein structures, etc. can be represented most naturally as graphs.
Relational data: Another area that has attracted considerable interest is clustering relational (network) data. Unlike the clustering of graph data, where the objective is to partition a collection of graphs into disjoint groups, the task here is to partition a large graph (i.e., network) into cohesive sub graphs based on their link structure and node attributes.

III. USE OF WEKA FOR IMPLEMENTATION OF K-MEANS ALGORITHM FOR FAULT DETECTION

To implement k-means algorithm WEKA tool can be used. The implementation of k-means algorithm will be used for predicting faults in the software modules. Fault prediction is a method used in software development life cycle to reduce the failure of software and takes place mostly during early planning to recognize fault-prone modules. Fault prediction not only raise the quality of monitoring during software development but also gives suggestions for suitable verification and validation approaches that eventually lead to improvement of efficiency and effectiveness of fault prediction. Nowadays, software development is mostly based on Object-Oriented (OO) paradigm. The quality of OO software can be best assessed by the use of software metrics. Many metrics have been proposed by researchers and practitioners to appraise the quality of software. Software fault prediction techniques require software metrics that can be collected with the help of automated tools and fault data belonging to previous software version or similar software project. Fault prediction models generally define one dependent variable and n independent variables. Dependent variable shows whether the module is fault-prone or not. Independent variables can be product or process metrics, collected from different studies focused on product metrics. Cyclomatic complexity and lines of code are some examples of method-level product metrics.

1) Performance evaluation parameters:
The following are the performance parameters that can be used for fault prediction.

- True Negative
- True Positive
- False Negative
- False Positive
- Precision
- Recall
- True positive rate
- False positive rate
- Area Under Curve(AUC)

TABLE I: CONFUSION MATRIX

	Non-Faulty	**Faulty**
Non-Faulty	True Negative (TN)	False Positive (FP)
Faulty	False Negative (FN)	True Positive (TP)

2) Data set:

Any data set can be input to the WEKA. Mostly open source datasets are used by the users.

3) Methodology:

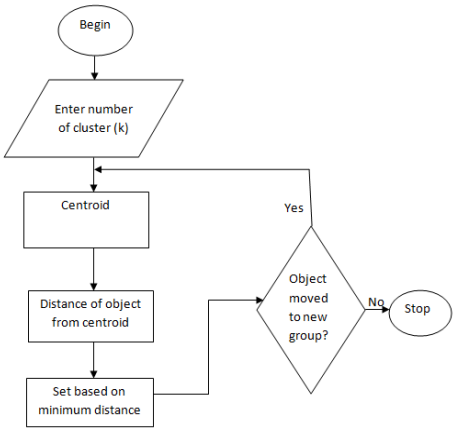

Fig3.K-means algorithm flowchart

4) Cluster Centroids:

The clusters are generated using K means clustering algorithm. Depending upon the lowest distance from centroid the cluster has been produced. The instance close to the centroid is assigned to the corresponding cluster. In this way the clusters are produced. Further, these centroid values are calculated using the Euclidean distance formulation as follow.
Euclidean distance between a point *X* (*X1*, *X2*, etc.) and a point *Y* (*Y1*, *Y2*, etc.)

$$D=\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

A new centroid is generated till a fresh cluster is produced from the new centroid. When the same cluster is generated from the new centroid as formed by prior centroid, the final cluster is formed and no any further iteration takes place. Also the finishing centroid value is obtained.

The following are the parameters that are used for the calculation of datasets.

TABLE IV: PARAMETERS USED

Name	Functionality	Value set
distanceFunction	The distance function is used to evaluate the instances of data set.	Euclidean Distance
dontReplaceMissingValues	Substitute omitted values worldwide through mean/mode	False
maxIteration	It sets the maximum number of iterations that will occur during the whole clustering/classification process.	500
numClusters	This option let us define the number of clusters needed in output.	2
preserveInstancesOrder	This option can preserve the order of instances of data set.	False
Seed	It defines the number of seed that will be used in clustering procedure.	10

III. CONCLUSION

The various developments, methods and improvements done over the years, are available for the different clustering techniques which are discussed in this paper. Many difficulties and challenges are faced by the users regarding clustering that are still obligatory to be recovered. The k-means algorithm is straightforward and well known algorithm which can be simply implemented using WEKA for fault detection.

REFERENCES

[1] Gantz,John F.,” The diverse and exploding digital universe”, “http://www .emc.com/collateral/analyst-reports/diverse-exploding-digital universe.pdf”
[2] Tukey,John Wilder,”Exploratory Data Analysis”.Addison-Wesley.Umeyama,S.,1988.An eigen decomposition approach

to weighted graph matching problems.IEEE Trans.Pattern An al. Machine Inte ll.10(5),695–703 .
[3] Duda ,R.,Hart,P.,Stork ,D. ,"Pattern Classification". John Wiley and Sons,New Yo r k.
[4]JSTOR, 2009. JSTO R.< http :// www. j s to r .o r g>.
[5] Frank, I ldiko E .,Todeschini,Roberto ,1994."Data Analysis Hand book ". Elsevier Science Inc ., pp . 227 – 228 .
[6] Dempster,A.P ., Laird , N.M. ,Rubin ,D.B. ,1977 ."Maximum likelihood from in complete data via the EM algorithm". J. Ro y. Statist.Soc. 3 9, 1 – 38 .
[7] Agrawal , Rakesh , Gehrke, Johannes , Gunopulos , Dimitrios , Raghavan , Prabhakar,1998."Automatic subspace clustering of high dimensional data for data mining applications".In : Proc . ACMSIG MOD, pp . 94 – 1 0 5.
[8] Jain,Anil K.,Dubes,Richard C . ,1988 ."Algorithms for Clustering Data". Prentice Hal l. Jain ,Anil K. , Flynn,P.,1996.Image segmentation using clustering.In :Advances in Image Understanding . IEEE Computer Society Press,pp.65 – 83
[9] Ball,G.,Hal l,D.,1965.ISO DATA,anovel method of data analysis and pattern classification.Technical report NTISAD 69 9616.Stanford Research Institute, Stanford ,CA.
[10] Forgy, E .W .,1965.Cluster analysis of multivariate data : Efficiency vs .interpretability of classifications .Biometrics 21,768– 769 .
[11] Dunn,J.C .,1973.A fuzzy relative of the ISO DATA process and its use in detecting compact well -separated clusters. J. Cybernet.3,32 –57 .
[12] Bezdek,J.C ., 1981."Pattern Recognition with Fuzzy Objective Function Algorithms" .PlenumPress .
[13] Jain ,A nil K.,Dubes , Richard C. ,1988 ."Algorithms for Clustering Data". Prentice H al l .
[14] Wallace , C. S., Boulton , D. M. , 1 9 6 8 ." An information measure for classification" .Comput . J. 1 1 , 1 8 5 – 19 5 .
[15] Figueiredo ,Mario ,Jai n ,Anil K .,2002 ."Unsupervised learning of finite mixture models". IEEE Trans. Pattern Anal . Machine Intel l .2 4 (3),381– 396 .
[16] Fisher , L .,van Ness , J., 1971 . "Admissible clustering procedures" .Biometrika .
[17] Fred ,A ., Jain , A. K., 2002 . "Data clustering using evidence accumulation" . I n: P roc. Internat . Conf. Pattern Recognition (I CP R) .
[18] Chapelle , O . , Scholkopf, B ., Z ie n , A. (E d s .), 2 0 0 6. "Semi - Supervised Learning". MIT Press , Cambridge , MA .
[19] Basu ,Sugato , Banerjee ,Ar indam, Mooney, Ray mo n d, 2 00 2 . "Semi- supervised Basu,Sugato,Banerjee" , Arindam,Mooney,Raymond, 2 00 2 . Semi- supervised
[20] Law , Martin , Topchy ,Alexander , Jai n , A.K ., 2 0 0 5 . "Model - based clustering with probabilistic constraints" .In : Proc. SIAM Conf. on Data Mining , p p . 64 1 – 6 4 5.
[21] Figueiredo , M. A. T. , Chang, D. S. , Murino , V. , 20 0 6 . "Clustering under prior knowledge with application to image segmentation" . Ad v. Neural Inform. Process . Systems 19 , 4 01 – 4 0 8.