ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Comparative Study of Clustering Algorithms: Filtered Clustering and K-Means Cluttering Algorithm Using WEKA

G.Thangaraju[1], J.Umarani[2], Dr.V.Poongodi[3]

Guest Lecturer, Department of Computer Science, Government Arts and Science College, Veppanthattai, India[1]

Assistant Professor, Department of Computer Applications, Thanthai Hans Roever College, Perambalur, India[2,3]

ABSTRACT: Clustering is an unsupervised learning problem which is used to determine the intrinsic grouping in a set of unlabeled data. Grouping of objects is done on the principle of maximizing the intra-class similarity and minimizing the inter-class similarity in such a way that the objects in the same group/cluster share some similar properties/traits. There is a wide range of algorithms available for clustering. This paper presents a comparative analysis of various clustering algorithms. In experiments, the effectiveness of algorithms is evaluated by comparing the results on 2 datasets from the UCI repository**.**

KEYWORDS: Data Mining, Clustering,Filtered clustering and K-means clustering, WEKA.

I. INTRODUCTION

Cluster

A Cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The process of grouping a set of physical or4 abstract objects into classes of similar objects is called ***clustering.***

Cluster Analysis

It has wide applications, including market or customer segmentation, pattern recognition, biological studies, spatial data analysis, Web document classification, and many others. Cluster analysis can be used as a stand-alone data mining tool to gain insight into the data distribution or can serve as a preprocessing step for other data mining applications operating on the detected clusters.

The quality of clustering can be assessed based on a dissimilarity of objects, which can computed for various types of data, including interval-scaled,binary,categorical,ordinal,and ratio-scaled variables, or combinations of these variable types. For nonmetric vector data, the cosine measure and the Tanimoto coefficient are often used in the assessment of similarity.

Clustering is a dynamic field of research in data mining. Many clustering have been developed. These can be categorized into portioned methods, hierarchical methods, density-based methods, grid based methods, model-based methods, methods for high-dimensional data (including frequent pattern-based methods. Some algorithms may belong to more than one category.

Partitioning Method

A Partitioning method first creates an initial set of k partitions, where parameter k is the number of partitions to construct. It then uses an iterative relocation technique that attempts to improve the partitioning methods include k-means, k-medoids, CLARANS, and their improvements.

Hierarchical Method

It creates the hierarchical decompositions of the given set of data objects. The method can be classified as being either agglomerative (bottom-up) or divisive (top-down), based on how the hierarchical decomposition is formed.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

To compensate for the rigidity of merge or split, the quality of hierarchical agglomeration can be improved by analyzing object linkages at performing micro clustering (that is grouping objects into "microclusters") and then operating on the micro clusters with otherClustering technique, such as iterative relocation.

Density-Based Method

A density-based method cluster objects based on the notion of density. It either grows clusters according to the density of neighbourhood objects (Such as DBSCAN) or according to some density function (such as DENCLUE). OPTICUS is a density based that generates as augmented ordering of the clustering structure of the data.

Grid-Based Method

A grid-based method first quantizes the object space into a finite number of cells that form a grid structure, and then performs clustering on the grid structure. STING is a typical example of grid-based method based on the statistical information stored in grid cells. WaveCluster and CLIQUE are two clustering algorithms that are both grid-based and density-based.

Model-Based Method

A model based method hypothesizes a model for each of the clusters and finds the best fit of the data to that model. Examples of model-0based clustering include the EM algorithm (which uses a mixture density model), Conceptual clustering such as (COBWEB), and neural network approaches (such as self-organizing feature maps).

Clustering high-dimensional data

Clustering high-dimensional data is of crucial importance, because in many advanced applications, data objects such as text documents and microarray data are high –dimensional in nature. There are three typical methods to handle high-dimensional data sets: dimension-growth subspace clustering represented by CLIQUE, dimension-reduction projected clustering, represented by PROCLUS, and frequent pattern-based clustering, represented by pCluster.

Constrained-based clustering method

It groups object based on application-dependent or user-specified constraints. For example, clustering with the existence of obstacle objects and clustering user-specified constraints, and semi-supervised clustering based on "weak" supervision.

Outlier detection and Analysis

One person's noise could be another person's signal. Outlier detection and analysis are very useful for deduction, customized marketing, medical analysis, and many other tasks. Computer based outlier analysis methods typically follow either a statistical distribution-based approach, a distance based approach, a density-based approach.

II. RELEATED WORK

In this Section, some of the projected clustering algorithms and the methodologies used in those algorithms to discover projected clusters are presented. In addition to projected clustering algorithms, two base algorithms such as k-means algorithm and fuzzy c-means algorithm that are used in our approach to discover projected clusters are also presented.

This study focuses only on projected clustering algorithms since projected clustering algorithms produce only disjoint clusters of high quality.

(Aggarwal et al 2005) introduces a algorithm named as ORiented projected CLUStering (ORCLUS). In their algorithm, projected cluster is defined as a set of eigenvectors together with a set of C clusters so that the data points in *C* are similar to each other in eigenvectors. ORCLUS is able to discover arbitrary oriented sub spaces and is able to scale to huge databases. It takes two inputs from the user i)*k* – number of clusters and ii) *l* – number of average dimensions to be included in each cluster. There are 3 steps in this algorithm i) assign ii) find vectors iii) Cluster merging.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

In the first step, the database is partitioned into specified number of clusters mentioned as per the input value by assigning each data point to its closest seed. The distance between the data point and the seed is calculated only in the sub space of eigenvectors. Iteratively the seeds are replaced by the centroids of the cluster which was just created.

In the second phase, sub space of *l* dimensions is calculated for each cluster. This is done by calculating the covariance matrix for each cluster, and by picking up the corresponding orthonormal eigenvectors with least Eigen values.

In the third phase, closest pairs of similar clusters are merged successively. This phase is terminated when all closed pairs are merged and *k* clusters are formed iteratively.

The main advantage of this method is avoidance of initial error since random seeding is reduced due to formation of large number of clusters in the beginning.

The main drawback of this algorithm is the request for 2 input parameters such as *k* – number of clusters and *l* – average dimensions in each cluster.

Input: Database containing n objects and k (Number of clusters) and *l* (number of dimensions)
Output: Formation of k clusters
Step 1: Assign: Selection of initial seeds and assignment of data points to seeds based on distance only in eigenvectors.
Step 2: Find Vectors: Ortho normal eigenvectors with least eigenvalues are picked. Covariance matrix for each cluster is formed. Sub space of *l* dimensions is identified.
Step 3: Similar clusters are merged and k clusters are formed iteratively.

Figure 2.1 Steps in ORCLUS

A Hierarchical approach with Automatic Relevant attribute selection for Projected Clustering was proposed by Yip et al (2004). The main advantage of HARP is its ability to determine automatically the relevant attributes of each cluster without requesting any input parameters from the user. The clustering quality of previous projected clustering algorithms depends mainly on input parameters such as number of clusters from the user.

A function is defined to measure the relevance of a dimension to a cluster. The relevance of a dimension is computed by computing local variance (variance within the cluster) and global variance (variance in the whole data set). The index value for a dimension is high if the local variance is extremely small. If the local variance is very high then the relevance index of a dimension is zero. The quality of the clusters is calculated as the sum of index values of all selected dimensions.

The main assumption made by HARP is that if two records are similar to each other, then, they have a high probability of belonging to the same cluster. At any time, a cluster selects dimensions with relevance, up to a certain level and allows similar clusters to be merged.

Main advantages of HARP are i) It provides a mechanism to select relevant dimensions automatically ii) it ensures that the selected dimensions have a guaranteed relevance to the cluster.

Disadvantage of HARP is that the assumption it made about two similar records always may belong to same cluster will not be accepted always.

Input: Database containing n objects
Output: Formation of clusters
Step 1: Local variance and global variances are computed.
Step 2: Relevance of a dimension is measured using a function by making use of local and global variances.
Step 3: Quality of a cluster is computed as the sum of index valuesof all selected dimensions.

Figure 2.2 Steps in HARP

Efficient Projected Clustering is proposed by Ng and Fu(2005) . It does not require any input parameter like number of clusters, average cluster dimensionality etc. It can handle clusters of irregular shapes. Produces best clustering results. It is scalable to large data sets.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

In this algorithm, projected clusters are treated as connected regions where densities are higher than the surrounding regions. A density estimation function f(x) is constructed. The data values that are near to x, only influences the density estimation function (Silverman 1986).

One dimensional histogram that was constructed on each dimension is used to build the density estimation function. An adaptive approach is applied on each dimension to discover dense regions where data objects aredensely located when projected at the dimension. The regions of histograms, where densities exceed some threshold determined by the mean value and standard deviation of the histogram, are located. Clustering regions of different densities are uncovered by adaptive method.

Once the dense regions for all dimensions are uncovered, then signature for a data point is computed by analyzing whether the data pointbelongs to dense region of many dimensions.

All data points with such signatures are combined to form potential clustering region. Projected clustering quality is evaluated in terms correctly discovered dimensions, correctly partitioned data objects etc.

Finally scalability of EPC is evaluated by varying number of data points, number of dimensions etc.

Input: Database containing n objects
Output: Formation of k clusters
Step 1: Assign: Selection of initial seeds and assignment of datapoints to seeds based on distance only in eigenvectors.
Step 2: Find Vectors: Ortho normal eigenvectors with least eigenvalues are picked. Covariance matrix for each cluster is formed. Sub space of *l* dimensions is identified.
Step 3: Similar clusters are merged and k clusters are formed iteratively.

Figure 2.3 Steps in EPC

HPSTREAM is a high dimensional projected data stream method proposed by Aggarwal et al (2004). This algorithm combines projected clustering with fading cluster structure. Data streams have attained much importance because of advances in hardware technology. Advances in hardware technology allow easy storage of numerous transactions in an automated way. Huge presence of data streams paves the way for lot of research (Berger and Rigoutsos 1991; Domingos and Hulten 2000; Guha et al 2000). HPStream takes 2 input parameters. 1. *k*– number of target clusters. 2. *l*– average dimensionality of projected clusters.

The historical and current data with a user defined fading factor are integrated with fading cluster structure. Clustering efficiency and quality have improved because of dynamic update of relevant dimensions and minimal radius for quality enhancement. The data stream consists of a set of records X1, X2 …Xk with time stamps T1, T2…Tk. A function f(t) at time t assigns weighttage to each data point. The function f(t) is also called as fading function.

The Fading cluster structure is defined as a data structure which isdesigned to capture key statistical characteristics of the clusters generatedduring the course of a data stream. The key characteristics of the underlying cluster are computed by the fading cluster structure. In order to weigh differentdimensions correctly, at the beginning of the clustering process, anormalization process is performed.

In order to compute radius of each dimension, this normalizationprocess is required. The domain of different attributes such as age, salary hasvast variances and ranges. The normalization process is mainly used equalizethe standard deviation in each dimension. A random sample data point ispicked to calculate the standard deviation i of each dimension i. The valueof i may changes because data streams may change from time to time.

The normalization factor is also recomputed at frequent intervals.

The statistics of fading structure needs to be changed whenever the value of i changes. For each data point, first to which cluster it has to be added is determined. The set of dimensions associated with each cluster is determined.

In order to know whether to include data point in the existing cluster or not is decided based on the natural limiting radius of the cluster. Data points which lie outside the natural boundary of a cluster create their own clusters. If it is necessary to create a cluster for a data point to be included, one existing cluster needs to be deleted.

HPSTREAM is highly scalable to large data sets and more dimensions. HPSTREAM outperforms CLUSTREAM in cluster purity.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Input: Database containing n objects and k (Number of clusters)
and*l* (average number of dimensions)
Output: Formation of k clusters
Step 1: Normalization process is performed to compute radius of each dimension. It equalizes standard deviation.
Step 2: A random data point is picked to calculate the standard Deviation i of each dimension i. Normalization factor is Re-computed at regular intervals.
Step 3: For each data point, first to which cluster it has to be added is determined. Associated dimensions are also determined

Figure 2.4 Steps in HPSTREAM

PCKA ALGORITHM

PCKA algorithm is proposed by Mohamed and Shergrui (2009) and it comprises of 3 steps

i. Attribute relevance analysis
ii. Outlier Handling and discovery of projected clusters. This algorithm makes use of partitioning concept and is able to produce low dimensional projected clusters which are embedded in high dimensional space.

PCKA discovers axis parallel clusters with following properties

i) Each projected cluster must be dense. Different clusters canhave different data points.Clusters allowed tohave common dimensions between them

iii) The data points within the clusters must be similar and theyshould be dissimilar to other data points in other clusters.

PCKA imposes no restrictions on number of data points as well asselected dimensions. There are 3 steps involved in PCKA algorithm: They are

i) Attribute relevance analysis: This is the first phase of .PCKAwhich is used to identify relevant dimensions for each cluster.

In a high dimensional space, some of the dimensions arealways not be relevant to any of the cluster. These dimensionsshould be identified and removed from the data set. Somedimensions of the data set only exhibit cluster structure.

Cluster structure simply means that some regions have higherdensity of points than their surrounding regions. Thesedimensions can be identified by calculating the dense regionsin each dimension. A sparseness degree for each onedimensional data point is calculated by measuring the varianceof its k-nearest neighbours. In order to know the characteristicsof each dimension, a probability density function is calculated.

Then MDL (Minimal Description Length) principle is used toidentify interesting sub spaces. This phase requires an inputparameter k, number of nearest neighbours of 1-d point. Finally, abinary matrix is formed which contains useful information aboutdense regions and their locations in the data set.

ii) Outlier Detection: Outliers are highly dissimilar andexceptional from other data points. In order to know thesimilarity between two given binary data points z1 and z2,four fundamental quantities are to be considered (Li 2006) asfollows a = | z1j = z2j = 1|

ii) b= |z1j=1 ^ z2j=0|

iii) c= |z1j=0 ^ z2j=1|

iv) d= |z1j=0 ^ z2j=0|

The similarities between binary values are measured using Jaccardcoefficient and is calculated using the formula

JC (zi, zj) = a / a+ b+ c

Jaccard coefficient has values between 0 (dissimilar) and 1 (mostsimilar). A data point is said to be outlier if the similarity between the binaryvalue of it and binary value of other data points are less than thresholds and, both are user defined parameters.

iii) Discovery of projected clusters: It is executed in 2 steps. Inthe first step, data points are clustered using k-meansalgorithm by measuring the distance between the data pointsrestricted to sub set of dimensions. In the second step relevantdimensions of the identified clusters are selected by makinguse of the properties of binary matrix.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Input: Database containing n objects and k (Number of clusters)
number of nearest neighbours, , , and .
Output: Formation of k clusters
Step 1: Attribute Relevance Analysis: A sparseness degree for eachOne dimensional point is calculated and binary matrixformed.
Step 2: Outlier Detection: Outliers are detected using JaccardCoefficient and user defined parameters, .
Step 3: Data points are clustered using k-means algorithm andrelevant dimensions are selected.

Figure 2.5 Steps in PCKA

III. CLUSTERING TECHNIQUES TO COMPARE

3.1 Filtered Clusterer

In mathematics, a filter is a special subset of a partially ordered set. For example, the power set of some set, partially ordered by set inclusion, is a filter. Let *X* be a topological space and *x* a point of *X*. A filter base *B* on *X* is said to cluster at *x* (or have *x* as a cluster point) if and only if each element of *B* has nonempty intersection with each neighbourhood of *x*.

- ✓ If a filter base *B* clusters at *x* and is finer than a filter base *C*, then *C* clusters at *x* too.
- ✓ Every limit of a filter base is also a cluster point of the base.
- ✓ A filter base *B* that has *x* as a cluster point may not converge to *x*. But there is a finer filter base that does. For example the filter base of finite intersections of sets of the sub base B U Nx.
- ✓ For a filter base *B*, the set ∩{cl(*B*0) : *B*0∈*B*} is the set of all cluster points of *B* (note: cl(*B*0) is the closure of *B*0). Assume that *X* is a complete lattice.
- ✓ The limit inferior of *B* is the infimum of the set of all cluster points of *B*.
- ✓ The limit superior of *B* is the supremum of the set of all cluster points of *B*.

B is a convergent filter base if and only if its limit inferior and limit superior agree; in this case, the value on which they agree is the limit of the filter base.

3.2 k-Mean Clustering

K-means clustering technique [24] is one of the simplest unsupervised learning techniques that aim to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean value. Initially, k centroids need to be chosen in the beginning. The next step is to take instances or points belonging to a data set and associate them to the nearest centers. After finding k new centroids, a new binding has to be done between the same data set points and the nearest new center. Process is repeated until no more changes are done. Finally, this algorithm aims at minimizing intra cluster distance (cost function also known as squared error function), automatically inter cluster distance will be maximized.

$$Cost_{Fun} = \sum_{i=1}^{k} \sum_{p \in C_i} ||p - m_i||^2$$

where, mi – mean of ithcluster, Ci- ithcluster and p – point representing the object.

IV. WEKA

WEKA (Waikato Environment for Knowledge Analysis) is an open source, platform independent and easy to use data mining tool issued under GNU General Public License. It comes with Graphical User Interface (GUI) and contains collection of data preprocessing and modeling techniques. Tools for data pre-processing, classification, regression, clustering, association rules and visualization as well as suited for new machine learning schemes are provided in the package. It is portable since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

4.1. User interfaces

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow as well as the command line interface (CLI). There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets. The *Explorer* interface features several panels providing access to the main components of the workbench:

The *Preprocess* panel has facilities for importing data from a database, a csv or an arff file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data from numeric to discrete, to remove missing instances, to appropriately choose missing values and converting csv file to arff and vice versa.

The Classify panel enables the user to apply classification and regression algorithms to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize errors. There are various type of classification algorithms like rule based, decision tree, naïve Bayesian, lazy, mi, misc etc. This paper make use of decision tree classification algorithms.

The Associate panel attempts to identify all important interrelationships between attributes in the data with the help of association learners like apriori, filtered associator, predictive apriori etc.

The *Cluster* panel gives access to the clustering techniques in Weka, e.g., the simple k-means, cobweb, DBSCAN, CLOPE algorithm to provide different kind of clustering's for different situations and usage of their results.

The *Select attributes* panel provides algorithms for identifying the most predictive attributes in a dataset.

The *Visualize* panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

V. METHODOLOGY & PERFORMANCE MEASURES

Clustering techniques discussed in section 3 have been compared with the help of WEKA. Steps followed in the analysis are:

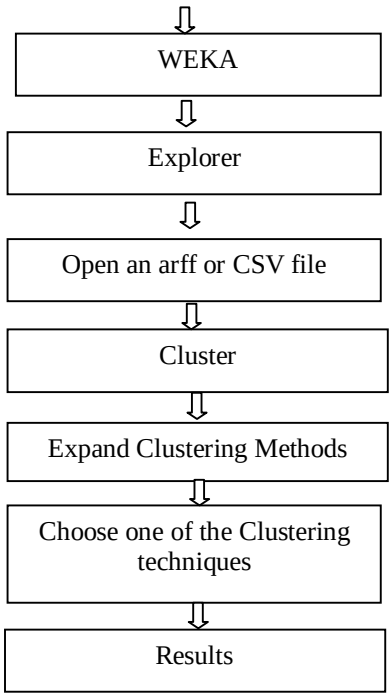

Figure:5.1. Clustering process used in WEKA

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

As shown in Fig3, data file supplied should be in arff or CSV form. Data File should not contain unique id attribute like names, roll nos., remove these attribute either before supplying it for classification or untick these attribute before classification in WEKA. Note that Weka also provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka ***Performance measure*** used to determine accuracy of clustered data is ***class to cluster evaluation***. A little about some important terms which are used in this measures is presented. These are:-

- ✓ True Clusterer (TC) – total number of elements belonging to clusters that were correctly predicted. These elements are verified using their classes i.e. TC= TC1 + TC2 + … TCn. Here n is the number of classes in the dataset and TCi is the number of elements of class Ci which belongs to correct/right cluster.
- ✓ N – Total number of instances which are clustered.

Accuracy: It determines the proportion of the total number of instances clustered to the instances which are correctly clustered.

$$\text{Accuracy} = \frac{TC}{N}$$

Execution time: Time taken to run the algorithm and produced the results.

VI. EXPERIMENTAL RESULTS

A comparative analysis of two clustering algorithms has been made using two datasets taken from the UCI machine learning repository. The details are summarized in Table 1.

Table 1. Dataset details

Datasets	#Instances	#Attributes
EPM-intermediate_grades	115	6
Sales_Transactions_Dataset_Weekly Data Set	811	53

Results are observed using two measures; accuracy and time, explained in section 4 using all the datasets mentioned in Table 1. Results have been shown in the Table 2, 3.

Table 2. EPM Data Set Analysis

Clustering Method	Accuracy (%)	Time Taken (in secs.)
Filtered Clusterer	80	0.08
k-Mean	80	0.09

Table 3. Sales Data Set Analysis

Clustering Method	Accuracy (%)	Time Taken (in secs.)
Filtered Clusterer	80	0.48
k-Mean	80	0.27

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

In the analysis, two different measures have been used for comparing various clustering algorithms. From the results obtained in the Tables 2, 3,it can be seen that K-mean performs best among all in most of cases. Clustering accuracy in K-mean is maximum and time taken in clustering is minimum.

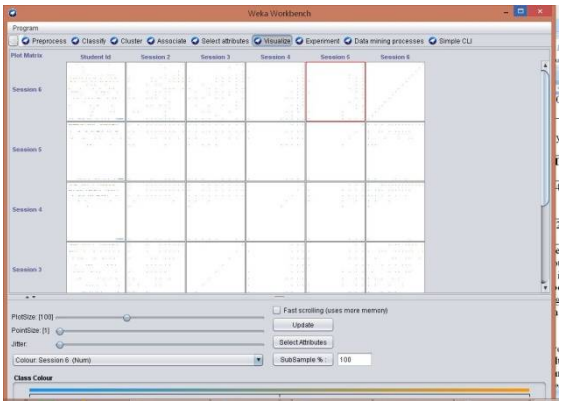

Figure 6.1: Visual image of k-means Data Set

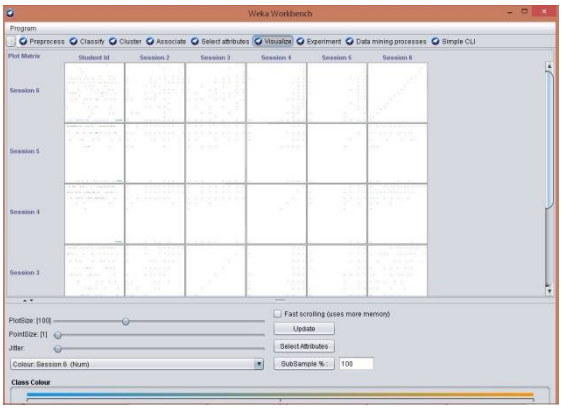

Figure 6.1: Visual image of k-mean for Sales Data Set

VII. CONCLUSIONS

Comparative analysis of twos clustering algorithms has been made. The results have been validated using two datasets taken from UCI repository and noticed that datasets are successfully clustered with a quite good accuracy. Few of the clustering techniques have better accuracy, others take less time, and many others have a trade-off between accuracy and time taken. Appropriate methods can be used according to their usage.

REFERENCES

1. Aggarwal, C.C., Han, J., Wang, J. and Yu, P.S. "A Framework for projected clustering of high dimensional data streams", Proceedings of the 30th VLDB conference, Toronto, Canada, 2004.
2. Yip, K.Y., Cheung, D.W. and Ng, M.K. "HARP: A practical projected clustering algorithm, IEEE transaction on Knowledge and Data Engineering, Vol 16, No 11, November 2004.
3. Ng. Eric KaKa and Fu. Ada Wai-chee, Wong. Raymond Ch-Wing., "Projective Clustering by Histograms", 2005.
4. Agrawal, R., Gehrke, J., Gunopulos, D. and Raghavan, P. "Automatic Subspace Clustering of High Dimensional Data. Data Mining and knowledge discovery", Springer Science + Business media, Inc. Manufactured in the Netherlands, Vol.11, pp.5-33, 2005.
5. Babcock, B., Babu, S., Datar, M., Motwani, R. and Widom, J. "Models and Issues in Data Stream Systems", ACM PODS Conference, 2002.
6. Han, J. and Kamber, M. "Data Mining: Concepts and Techniques", 2nd edition, Morgan Kaufmann publishers, San Francisco, CA, 2006.
7. Li, T. "A unified view on clustering binary data", Machine Learning, Vol. 62, No.3, pp 199-215, 2006.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

8. Mohamed, B. and Shergrui, W. "Mining Projected Clusters in High-Dimensional Spaces". IEEE Trans. on Knowledge and Data Engineering, Vol. 21, No. 4, April 2009.
9. Moise, G., Sander, J. and Ester, M. "P3C: A Robust Projected Clustering Algorithm", 2006.
10. Yip, K.Y., Cheung, D.W. and Ng, M.K., "On discovery of extremely low dimensional clusters using Semi Supervised Projected Clustering", Proceedings, pg 329-340, 21st International Conference on Data Engineering, 2005 (ICDE 2005).
11. Yiu, M.L. and Mamoulis, N. "Iterative Projected Clustering By Subspace Mining". IEEE Transactions on Knowledge and Data Engineering, Vol. 17, No. 2, pp. 1-14, 2005.
12. Yu, L.T.H., Chung, F.L. and Chan, S.C.F. "Emerging Pattern based Projected Clustering for Gene Expression Data", Proceedings of European workshop on Data Mining and Text Mining for Bio Informatics , held in conjunction with ECML / PKDD, 2004.