Clustering methods for Big data analysis

Keshav Sanse, Meena Sharma

Abstract— **Today's age is the age of data. Nowadays the data is being produced at a tremendous rate. In order to make use of this large-scale data and extract useful knowledge, researchers in machine learning and data mining communities are faced with numerous challenges. To manage and analyze such a big data in a specified time is the main challenge today. Clustering helps to visually analyze the data and also assists in decision making. Clustering is widely used in variety of applications like marketing, insurance, surveillance, fraud detection and scientific discovery to extract useful information. This paper contains an overview of the various clustering methods along with their classification, general working and also provides a comparison (from a theoretical perspective) among them. The paper presents the properties, advantages and drawbacks of the various clustering methods.**

Index Terms—**Cluster, clustering methods, outlier.**

I. INTRODUCTION

Today the massive data explosion is the result of a dramatic increase in the devices located at the periphery of the network including embedded sensors, smart phones and tablet computers. These large volumes of data sets are produced by the employees of the companies, social networking sites and different machines (cloud storage, meters, CC TV cameras etc). These large datasets are stored in the data centers. The problems associated with these large data sets or big data are their storage and management, retrieval and analysis. This paper is mainly related to the problem of analyzing such big data in a tolerable time. One way to overcome this problem is to have big data clustered in a compact format that will still be an informative version of the entire data. Clustering is a technique used for exploratory data analysis. It helps to analyze large volumes of data visually thus assists in making quick decisions.

Clustering is an unsupervised technique used to classify large datasets in to correlative groups. No predefined class label exists for the data points or instances. Clustering groups data instances into subsets in such a manner that similar instances are grouped together, while different instances belong to different groups and the groups are called as clusters.

Keshav Sanse, Department of Computer engineering, Institute of Engineering and Technology, DAVV university Indore, India

Meena Sharma, Department of Computer engineering, Institute of Engineering and Technology, DAVV university Indore, India

Thus, clustering is a division of data into groups of similar objects. Clustering can be classified into the following categories:

A. Partitioning clustering
B. Hierarchical clustering
C. Density based clustering
D. Model based clustering
E. Grid based clustering

In this paper we are investigating about big data clustering techniques. The clustering algorithms are compared using the following factors-

1. Dataset size- It means the volume or the size of the dataset; it can be small, medium or large.
2. Dataset type- It means the type of the attributes in the dataset. It can be numeric, categorical, mixed etc.
3. Cluster shape- A good clustering algorithm should be able to produce clusters of arbitrary shapes.
4. Time complexity- Time for obtaining the final clusters. A successful algorithm should take lesser time to cluster a large volume of data.
5. Handle outlier- Outliers are the points containing false information that make it difficult for an algorithm to cluster the data into the suitable or true cluster. Hence they should be properly handled or removed.

The paper is organized in the following manner: Section II is Related study; concluding in brief the views of researchers on clustering techniques as given in their respective research papers. Section III contains the description of various clustering methods. Section IV is Results of analysis it shows a table comparing the clustering algorithms on different factors. The paper is finally concluded in the section V.

II. RELATED STUDY

A number of surveys on clustering are present in the literature. Some researchers have proposed new algorithms for clustering. Other researchers have improved the existing clustering algorithms overcoming the drawbacks of the algorithm while some have performed a comparative study of the various clustering algorithms. Below is given a brief description of some of the previous studies on clustering methods-

- Reference [8] has compared the four clustering algorithms k-means, SOM, EM, hierarchical on some test dataset and extracted some conclusion on the quality, accuracy and performance of the four algorithms.

ISSN: 2278 – 1323

- Reference [9] has classified the clustering algorithms in to four types, shown their pros and cons and compared them on various factors.
- Reference [13] has described the various limitations of the k-means algorithm and the different techniques used to remove them.
- Reference [3] has given detailed description of the different clustering algorithms and also discussed about the scalability, dimensionality reduction and other general algorithmic issues.
- Reference [12] has proposed a modified k-means algorithm using three normalisation techniques and outlier detection technique and compared the performance with the traditional algorithm on UCI dataset repository.
- Reference [11] has compared six different clustering algorithms using three different dataset keeping in mind the size of dataset, number of clusters time taken to build clusters. Weka tool is used for comparing the performance.

III. CLUSTERING METHODS

A. *Partitioning method*

The partitioning based method divides data objects into a number of partitions (clusters). In this method, data objects are divided into non-overlapping subsets (clusters) such that all data objects into same clusters are closer to center mean values. In this method, all clusters are determined promptly. Partitioning methods relocate instances by moving them from one cluster to another, starting from an initial partitioning. In this method convergence is local and the globally optimal solution cannot be guaranteed.

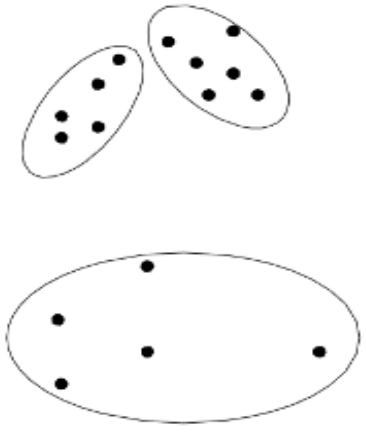

Figure shows partitioning clustering

Such methods typically require that the number of clusters should be pre-set by the user. This method minimizes a given clustering criterion by iteratively relocating data points between clusters until a (locally) optimal partition is attained. K-mean and K-medoids are examples of partitioning based methods [2]. The clusters in this method should fulfill these two requirements:-

(1) Each group or cluster must contain at least one object.
(2) Each object must belong to exactly one group or cluster.

1) K-mean

K-means clustering aims to partition n objects in to k clusters in which each observation belongs to the cluster with the nearest mean. Every cluster is having a cluster head or centroid. The centroid of a cluster is equal to the mean of all points in that cluster. The number of clusters is randomly chosen by user. K-mean algorithm proceeds by iteratively allocating points to the cluster with the closest centroid. The 'Closeness' is measured using Euclidean distance.

Algorithm-

1. Select K initial centroids
2. Repeat
   i) For each point, find its closest centroid and assign that point to the centroid. This results in the formation of K clusters
   ii) Re-compute centroid for each cluster

   Until the centroids do not change

Convergence criterion function for k-means algorithm is-

$$E = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - m_i \right\|^2$$

x_{ij} is the j^{th} data point of i^{th} cluster

m_i is the center of i^{th} cluster

n_i is the number of data points of i^{th} cluster

The above function is SSE (sum of squares error), for each point, the error is the distance to the nearest cluster. Clustering solution with less SSE is more optimal as compared to other clustering solutions.

K-means is a simple algorithm. The use of the K-means algorithm is limited to numeric attributes. There exists an algorithm called K-prototypes algorithm, which is based on the K-means algorithm but removes numeric data limitations while preserving its efficiency. The algorithm groups objects with numeric and categorical attributes in a way similar to the K-means algorithm. The similarity measure on numeric attributes is the squared Euclidean distance while the similarity measure on the categorical attributes is the number of mismatches between data objects and the cluster prototypes.

Advantages-

1. Easy to understand and implement.
2. Produce more dense clusters than the hierarchical method especially when clusters are spherical.
3. For large number of variables, K-means algorithm may be faster than hierarchical clustering, when k is small.
4. Efficient in processing large datasets.

Drawbacks-

1. Poor at handling noisy data and outliers.
2. Works only on numeric data.

ISSN: 2278 – 1323

3. Empty cluster generation problem.
4. Random initial cluster center problem.
5. Not suitable for non-spherical clusters.
6. User has to provide the value of k.

2) K-medoids

This algorithm is very similar to the K-means algorithm. It varies from the k-means algorithm mainly in its representation of the different groups or clusters. Each cluster is represented by the most centric object in the cluster, rather than by the implicit mean that may not belong to the cluster. For every point in the cluster: add up all distances to the other points in the cluster. Point in the cluster for which this distance is the smallest, becomes the medoid or centroid.

Both k-means and k-medoids require the user to specify K, the number of clusters. Two well-known types of k-medoids clustering [3] are the PAM (Partitioning Around Medoids) and CLARA (Clustering Large Applications)

Algorithm-

- Starts from an initial set of medoids and clusters are generated by points which are close to respective medoids.
- The algorithm iteratively replaces one of the medoids by one of the non-medoids if it improves the total distance of the resulting clustering.

The K-medoids method [10] is more robust than the K-means algorithm. In the presence of noise and outliers, a medoid is less influenced by outliers or other extreme values than a mean. However, its processing is more costly than the K-means method.

B. *Hierarchical method*

As the name suggests, this method builds clusters in a hierarchical order i.e. it forms nested clusters organised in a hierarchical tree. It forms clusters by recursively or iteratively partitioning the instances in either a top-down or bottom-up fashion.

Hierarchical clustering method is of two types [1]-

1. Agglomerative- this is a bottom up approach. In this approach initially each object is considered as a separate individual cluster. It then merges two or more suitable clusters to form new clusters. This merging of clusters is done recursively until a desired cluster structure or a stopping criterion (desired number of clusters k) is reached.
2. Divisive- this is top down approach. In this approach, initially the entire dataset is considered as one cluster. The cluster is then divided in to sub-clusters, which in turn are successively divided into more sub-clusters. This process is repeated until the stopping criterion (desired number of clusters k) is met.

Hierarchical method makes use of various criteria for performing the merging or splitting of clusters. The criteria are based on similarity measure or measure of cluster proximity. There are three measures of cluster proximity-

1. Single-link- It is also known as the minimum method or the nearest neighbor method. In this, the distance between two clusters should be the minimum distance from any member of one cluster to any member of the other cluster.

Drawback- "chaining effect": A few points that form a bridge between two clusters cause the single-link method to unify these two clusters into a single cluster.

2. Complete-link- It is also known as the diameter, the maximum method or the furthest neighbor method. In this, the distance between two clusters should be the longest distance from any member of one cluster to any member of the other cluster.
3. Average-link- It is also known as the minimum variance method. Here, the distance between two clusters is equal to the average distance from any member of one cluster to any member of the other cluster.

Drawback- forces elongated clusters to split and make portions of neighboring elongated clusters to merge.

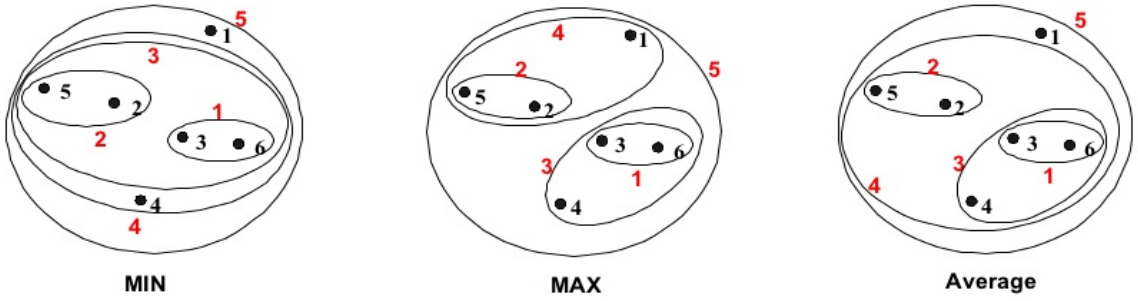

Figure shows hierarchical clustering using single (min), complete (max) and average link

Hierarchical clustering is represented by a two dimensional diagram known as dendrogram which explains the merges or splits made at each successive stage of analysis. In a dendrogram each data object is represented by a leaf node. It shows the similarity levels at which groupings change. A cluster is obtained by cutting the dendrogram at the desired similarity level.

BIRCH, CURE and Chameleon are examples of hierarchical clustering algorithms.

Advantages-

1. It is more versatile
2. Less sensitive to noise and outliers.
3. Any number of clusters can be obtained by cutting the dendrogram at desired level. It allows different

ISSN: 2278 – 1323

users to choose different partitions according to the desired similarity level.
4. Applicable to any attribute type.

Drawbacks-

1. If an operation (merge or split) is performed, it cannot be undone i.e. no backtracking is possible.
2. Inability to scale well.

BIRCH (Balance Iterative Reducing Clustering) [4] is the first clustering algorithm which removes the noisy data or outliers. This algorithm is also called as hybrid clustering. It overcomes the difficulties of hierarchical method: scalability and no backtracking. It makes full utilization of the memory and minimizes the I/O cost.

CURE (Clustering Using Representative) [5] uses sample point variant as the cluster representative rather than every point in the cluster. It identifies a set of well scattered points, representative of a potential cluster's shape. It scales/shrinks the set by a factor α to form semi-centroids and merges them in successive iterations. It is capable of finding clusters of arbitrary shapes.

C. *Density based method*

Density-based clustering method is based on the concepts of density, connectivity and boundary. This method forms clusters based on the density of data points in a region and continue growing a given cluster as long as the density (number of objects or data points) in the neighborhood is exceeding some threshold. Therefore, each data instance in the cluster the neighborhood of a given radius has to contain at least a minimum number of objects.

This method builds clusters of arbitrary shape since the cluster grows in any direction the density leads to. As this method forms clusters based on density of data points, it naturally eliminates the outliers or noisy data points. DBSCAN [6], OPTICS [7] and DENCLUE are examples of density based algorithms.

1) DBSCAN

This algorithm forms clusters using two parameters:

- Eps ε: Maximum radius of the neighborhood
- MinPts: Minimum number of points in an Eps neighborhood of that point

The algorithm forms clusters by searching the ε-neighborhood of each object in the dataset and checks if it contains more than the minimum number of objects. If the ε-neighborhood of any point p contain more than MinPts, new cluster with p as a core object is formed. DBSCAN then iteratively collects directly density-reachable objects from these core points, which involve the merge of a few density-reachable clusters. This process terminates when no new point can be added to any cluster.

A point p is directly density-reachable from a point q w.r.t. Eps and MinPts if

– p belongs to NEps(q)
– Core point condition:
|NEps(q)| >= MinPts

A point p is density-reachable from a point q w.r.t. Eps, MinPts if there is a chain of points q = p1, p2, ..., pn = p such that pi+1 is directly density reachable from pi.

A point p is density-connected to a point q w.r.t. Eps, MinPts if there is a point o such that both p and q are density-reachable from o w.r.t. Eps and MinPts

Cluster = set of density-connected points

Core point: A point is a core point if it has more than a specified number of points (MinPts) within Eps. These points are at the interior of a cluster.

Border point: A border point has fewer than MinPts within Eps, but is in the neighborhood of a core point. These points are at the outer surface of a cluster.

Noise point: A noise point is any point that is not a core point or a border point. These points are not part of any cluster.

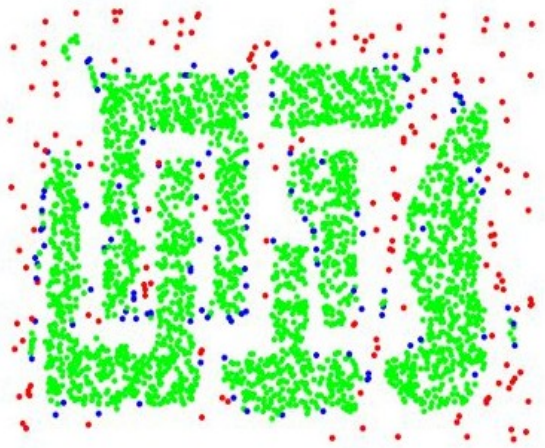

Figure shows DBSCAN: core points (in green), border points (in blue) and noise points (in red) Eps=10, MinPts=4

Algorithm-

Repeat
– Select a point p
– If p is core point then,
Retrieve and remove all points density-reachable from p w.r.t. Eps and MinPts;
Output them as a cluster
Until all points have been processed

OPTICS (Ordering Points to Identify Clustering Structure) is another algorithm similar to DBSCAN for finding density based clusters in spatial data. It addresses one of DBSCAN'S major weaknesses i.e. of detecting meaningful clusters in data of varying density. DENCLUE is yet another algorithm which forms clusters using the concept of density attracted regions.

Advantages-

1. Resistant to outliers.
2. Does not require the number of clusters.
3. Forms clusters of arbitrary shapes.
4. Insensitive to ordering of data objects

Drawbacks-

1. Unsuitable for high-dimensional datasets due to the curse of dimensionality phenomenon.

ISSN: 2278 – 1323

2. Its quality depends upon the threshold set.

D. *Model based* [10]

Model based clustering method optimizes the fit between the given data and some (predefined) mathematical model. It assumes that the data were generated by a model or by a mixture of underlying probability distributions and tries to recover the original model from the data. The model that we recover from the data then defines clusters and assigns objects to clusters. It leads to a way of automatically determining the number of clusters based on standard statistics taking noise (outliers) into account and thus yielding a robust clustering method. MLE (maximum likelihood estimation) criterion is used in model-based clustering method to find the parameter inside the model. The two major approaches based on the model-based clustering are: statistical and neural network approaches. EM (which uses a mixture density model), COBWEB (conceptual clustering) and neural network approaches (such as self-organizing feature maps) are examples of model based clustering methods.

1) EM (Expected Maximization)

EM finds the maximum-likelihood (ML) estimates for the parameters of the data model. The model parameters estimated by EM should be ML in the sense that they maximize the likelihood of all of the observed data. EM can decide how many clusters to generate. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step. EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters.

2) Self-organizing map (SOM)

SOM constructs a single-layered network. The learning process takes place in a "winner-takes-all" fashion: The prototype neurons compete for the current instance. The winner is the neuron whose weight vector is closest to the instance currently presented. The winner and its neighbors learn by having their weights adjusted. It is useful for visualizing high-dimensional data in 2D or 3D space. However, it is sensitive to the initial selection of weight vector, as well as to its different parameters, such as the learning rate and neighborhood radius.

Advantages-

1. Robust to noisy data or outlier.
2. Fast processing speed.
3. It decides the number of clusters to generate.

Drawbacks-

1. Complex in nature.

E. *Grid based* [9]

The grid based clustering uses a multi resolution grid data structure. It is used for building clusters in a large multidimensional space wherein clusters are regarded as denser regions than their surroundings. This method partition the space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. It differs from the conventional clustering algorithms in that it is concerned not with the data points but with the value space that surrounds the data points. The accumulated grid-data make grid-based clustering techniques independent of the number of data objects that employ a uniform grid to collect regional statistical data, and then perform the clustering on the grid, instead of the database directly. Grid based methods help in expressing the data at varied level of detail based on all the attributes that have been selected as dimensional attributes. In this approach representation of cluster data is done in a more meaningful manner. A typical grid-based clustering algorithm consists of the following five basic steps:

1. Creating the grid structure, i.e., partitioning the data space into a finite number of cells.
2. Calculating the cell density for each cell.
3. Sorting of the cells according to their densities.
4. Identifying cluster centers.
5. Traversal of neighbor cells

STING (Statistical Information Grid based) [14] and Wave Cluster are examples of grid based clustering. The quality of clustering produced by this method is directly related to the granularity of the bottom most layers, approaching the result of DBSCAN as granularity reaches zero. It explores statistical information stored in grid cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure: each cell at high level is partitioned to form a number of cells at the next lower level. Statistical information regarding the attributes in each grid cell is pre-computed and stored. CLIQUE was the first algorithm proposed for dimension –growth subspace clustering in high dimensional space. Wave Cluster does not require users to give the number of clusters applicable to low dimensional space. It uses a wavelet transformation to transform the original feature space resulting in a transformed space where the natural clusters in the data become distinguishable.

Advantages-

1. Fast processing time.
2. Independent of the number of data objects.

Drawbacks-

1. Depends only on the number of cells in each dimension in the quantized space.

IV. RESULTS

Results of analysis are given in the table I.

ISSN: 2278 – 1323

Table I shows result
(I=no. of iterations, k=no. of clusters, n=no. of objects)

Method name	**Algorithm**	**Time Complexity**	**Dataset size**	**Dataset type**	**Cluster shape**	**Handle outlier**
Partitioning	K-means	O(Ikn)	Huge	Numeric	Spherical	No
	k-medoids	$O(n^2I)$	Small	Categorical	Spherical	Yes
	k-prototype	O(n)	Small	Numeric & categorical	Spherical	No
Hierarchical	BIRCH	O(n)	Huge	Numeric	Spherical	Yes
	CURE	$O(n^2 \log n)$	Huge	Numeric	Arbitrary	Yes
	CHAMELEON	$O(n^2)$	Huge	All type of data	Arbitrary	Yes
Density based	DBSCAN	O(n log n)	Huge	Numeric	Arbitrary	Yes
	OPTICS	O(n log n)	Huge	Numeric	Arbitrary	Yes
	DENCLUE	O(n log n)	Huge	Numeric	Arbitrary	Yes
Grid based	STING	O(n)	Huge	Special	Arbitrary	Yes
	Wave Cluster	O(n)	Huge	Special	Arbitrary	Yes
Model based	EM	O(n)	Huge	Special	Spherical	No
	SOM	$O(n^2m)$	Small	Multivariate	Spherical	No

ISSN: 2278 – 1323

V. CONCLUSION

In this paper we studied the different clustering methods and the algorithms based on these methods. Clustering is widely used in a number of applications. Each clustering method is having its own pros and cons. From the table it is clear that none of the clustering algorithms discussed, performs well for all the governing factors.

K-mean (partitioning based) is the simplest of all the algorithms. But its use is restricted to numeric data values only. The performance of the k-mean algorithm increases with the increases as the number of clusters increase. Hierarchical method forms nested clusters by splitting or merging of data points. In this method no backtracking is allowed.

Density based method is designed for building clusters of arbitrary shapes. It builds clusters automatically i.e. no need to mention the number of clusters and naturally removes outliers.

Grid based method mainly concentrates on spatial data. EM algorithm provides excellent performance with respect to the cluster quality, excluding for high-dimensional data.

To use the appropriate clustering method solely depends upon our requirement, the application involved and the other governing factors as mentioned in the comparison table.

REFERENCES

[1] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta "A Comparative Study of Various Clustering Algorithms in Data Mining", 2012

[2] Jiawei Han and Micheline Kamber, Jian Pei, B Data Mining: Concepts and Techniques, 3rd Edition, 2007

[3] Pavel Berkhin, "Survey of Clustering Data Mining Techniques" 2002

[4] Tian Zhang, Raghu Ramakrishnan, Miron Livny "BIRCH: An Efficient Data Clustering Method for Very Large Databases"

[5] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim "An Efficient Clustering Algorithm for Large Databases"

[6] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise" in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996

[7] Mihael Ankerst, Markus M. Breunig, Hans-Peter Kriegel, Jörg Sander, "OPTICS: Ordering Points To Identify the Clustering Structure" Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999

[8] Osama Abu Abbas "Comparisons between data clustering algorithms"

[9] Preeti Baser, Dr. Jatinderkumar R. Saini "A Comparative Analysis of Various Clustering Techniques used for Very Large Datasets"

[10] Lior Rokach, Oded Maimon Chapter 15 CLUSTERING METHODS "DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK"

[11] Garima Sehgal, Dr. Kanwal Garg "Comparison of Various Clustering Algorithms" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014, 3074-3076

[12] Vaishali R. Patel1 and Rupa G. Mehta "Impact of Outlier Removal and Normalization Approach in Modified k-Means Clustering Algorithm" IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 2, September 2011

[13] Kehar Singh, Dimple Malik and Naveen Sharma "Evolving limitations in K-means algorithm in data mining and their removal" IJCEM International Journal of Computational Engineering & Management, Vol. 12, April 2011,

[14] Suman and Mrs.Pooja Mittal "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set Using the weak tool" IJETTCS.

Keshav Sanse is currently pursuing M.E. (software engineering) from IET DAVV, Indore. He is a teaching assistant in the institute. He has completed B.E. (computer science) from CITM, Indore.

Meena Sharma is teaching as a professor in IET DAVV, Indore. She has 14 years of teaching experience and 5 years of industry experience. Her educational qualification includes B.E. (computer engineering, 1992), M.Tech (computer science, 2004) and Ph.D (computer engineering, 2012).

ISSN: 2278 – 1323