A Survey on clustering

Current status and challenging issues

Rama.B[1]
Department of Informatics
Kakatiya University, Warangal

Jayashree.P[2]
Department of CSE
JITS, Karimnagar

Salim Jiwani[3]
Department of CSE
VCE, Warangal

***Abstract*—Clustering is the art of subset from a dataset. It helps in identifying the hidden information and arranging data into its logical group based on an attribute or a set of attributes. The intent of this paper is to explore a variety of clustering methods and brief their working styles so that researches can have a partial view of methods discussed. This work presents gains and pitfalls and also associated worst case complexities of each clustering method to some extent. The techniques discussed here are just a snap shot of clustering algorithms. Currently model based algorithms are been used to improve efficiency of clustering algorithms. This paper would be helpful in devising the choice of algorithm for such a purpose.**

***Key words-Data Mining; cluster; Survey; Complexities*.**

I. INTRODUCTION

In the area of intelligent data analysis cluster analysis is an unsupervised learning methodology constituting a cornerstone. It is used for exploration of intra and inner cluster relationships among a collection of patterns organized as clusters. The attributes (data) may be categorical, continuous or binary. Cluster analysis is a difficult problem hence devising a well tuned clustering technique for a given clustering problem is required. Moreover, it is well known that no clustering method can adequately handle all sorts of cluster structures and input data. It is not uncommon to try to find noisy values (outliers) and eliminate them by a preprocessing step.

The reader should be cautioned that a single article couldn't be a comprehensive review of all learning algorithms. Rather, our goal is to provide a representative sample of the research in each of the cluster learning technique.

Route map for the paper

Partitioning algorithms are described in section 2. Hierarchical algorithms are covered in section 3. The section 4 explains the density-based algorithms. The section 5 describes the grid-based methods. The model-based algorithms are covered in section 6, while, recent advances in clustering techniques, such as ensembles of clustering algorithms, are described in section 7. The final section concludes this work.

II. PARTITIONING METHODS

Partitioning algorithms tries to explore the subset in the dataset at a time. They may also be used in a top down approach. The methods in this approach are K-Means, Farthest First Traversal k-center (FFT) algorithm, K-Medoids (PAM), CLARA, CLARANS, Fuzzy K-Means, K-Modes, Fuzzy K-Modes, squeezer, K-prototypes, COOLCAT, etc. few methods are explained below.

Partitioning method works by dividing the data into subset or partition based on some evaluation criteria. Major divisions in this type are centroid based and medoid based algorithms. The centroid algorithms represent each cluster by using the gravity centre of the instances. The medoid algorithms represent each cluster by means of the instances closest to the gravity center. The most well-known partitioning algorithm is the k-means [2]. The k-means method partitions the batch of data i.e. entire data set into k subsets such that all points in a given subset are closest to the same centre for an attribute. The operation is iterated until there is no change in the gravity centers. The effectiveness of this method as well as of others relies heavily on the objective function used in measuring the distance between instances. The difficulty is in finding a distance measure that works well with all types of data. There are several approaches to define the distance between instances [2]. This algorithm is efficient in processing large data sets, often terminates at a local optimum, the clusters have spherical shapes and sensitive to noise [3].

The k-modes algorithm [1] is applicable for categorical attributes making use of simple matching coefficient measure recent. The k-prototypes algorithm, integrates the k-means and k-modes algorithms to allow for clustering instances described by mixed attributes.

Fuzzy clustering extends this notion to associate each pattern with every cluster using a membership function; here a data point may belong to more than one cluster producing non disjoint clusters. One widely used algorithm is the Fuzzy C-Means (FCM) algorithm [2], which is based on k-means. FCM attempts to find the most characteristic point in each cluster, which can be considered as the "center" of the cluster and, then, the grade of membership for each instance in the clusters. The strategy in this algorithm is to start with initial guesses for the mixture model parameters. These values are then used to calculate the cluster probabilities for each instance. These probabilities are in turn used to re-estimate the parameters, and the process is repeated. Other soft clustering algorithms have been developed and most of them are based on the Expectation-Maximization (EM) algorithm [4]. They assume an underlying probability model with parameters that describe the probability that an instance belongs to a certain cluster.

The drawback of such algorithms is that they tend to be computationally expensive and over fitting. In this context, one

possible solution is to adopt a fully Bayesian approach, in which every parameter has a prior probability distribution.

Partition clustering methods are applicable on single attribute provided user enters number of clusters and stopping criteria; they are order sensitive and cannot handle noisy data. They often tend to produce spherical shaped clusters only.

III. HIERARCHICAL CLUSTERING

The hierarchical methods group data instances into a tree of clusters (dendrogram). There are two major methods under this category. One is the agglomerative (bottom-up) method and divisive (top-down) method. These methods terminate quickly.

The advantages of Agglomerative and divisive methods are:

- Does not require the number of clusters to be known in advance,
- Computes a complete hierarchy of clusters,
- Good result visualizations are integrated into the methods,

However, the methods are not adaptive after a splitting or merging decision is made. Hierarchical clustering techniques use various criteria to decide "locally" at each step which clusters should be joined or split. For agglomerative hierarchical techniques, the criterion is typically to merge the "closest" pair of clusters, where "close" is defined by a specified measure of cluster proximity. There are three definitions of the closeness between two clusters: single-link, complete-link and average-link.

- The minimum distance between elements of each cluster (also called single-linkage clustering), good at handling non-elliptical shapes, but is sensitive to noise and outliers.
- The maximum distance between elements of each cluster (also called complete linkage clustering) and is less susceptible to noise and outliers, has trouble with convex shapes.
- The mean distance between elements of each cluster (also called average linkage clustering.

Some of the hierarchical clustering algorithms are: Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Clustering using representatives (CURE) and CHAMELEON, ROCK used with categorical data.

BIRCH [7] uses a hierarchical data structure called CF-tree for partitioning the incoming data points in an incremental and dynamic way (online training). It can typically find a good clustering with a single scan of the data and improve the quality further with a few additional scans. Noise can be handled effectively. It is reasonably fast and intelligent alternative to other clustering algorithms which are order-sensitive. Bubble and Bubble-FM [7] clustering algorithms are extensions of BIRCH to handle categorical attributes.

CURE, uses a constant number of representative points (parameters 'c') to represent a cluster. Inter cluster distance is measured as the closest pair of the representative points belonging to different clusters [7]. It takes a random sample to find clusters of arbitrary shapes (e.g. ellipsoidal, spiral, cylindrical, non-convex) and sizes, as it represents each cluster via multiple representative points for small data sets.

CHAMELEON [6] finds the clusters in the data set by using a two-phase algorithm. In the first phase it generates a k-nearest neighbor graph (using graph-partitioning algorithm) [6] that contains links only between a point and its k-nearest neighbors. During the second phase, it uses an agglomerative hierarchical clustering algorithm to find the genuine clusters by repeatedly combining together these sub-clusters. In this method no cluster can contain less than a user specific number of instances.

ROCK [9], is another clustering algorithm for categorical data using the Jaccard coefficient to measure similarity. The input is a set S of n sampled points to be clustered (that are drawn randomly from the original data set), and the number of desired clusters k. It samples the data set in the same manner as CURE.

A novel incremental hierarchical clustering algorithm (GRIN) for numerical data sets (based on gravity theory in physics). The algorithm is insensitive to the distribution of dataset.

Hierarchical clustering is mostly applicable on sample dataset or random sample. These algorithms are sensitive to order of presentation of data. They handle outliers and noisy data better than partition clustering methods. They also produce arbitrary shaped clusters representing data appropriately in the clusters. The user has to still provide number of clusters and stopping criteria.

IV. DENSITY-BASED CLUSTERING

Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well known density-based clustering algorithms is the DBSCAN [7]. DBSCAN separates data points into three classes:

- Core points: These are points that are at the interior of a cluster.
- Border points: A border point is a point that is not a core point, but it falls within the neighborhood of a core point.
- Noise points: A noise point is any point that is not a core point or a border point.

To find a cluster, DBSCAN starts with an arbitrary instance (p) in data set (D) and retrieves all instances of D with respect to Eps and Min Pts. The algorithm makes use of a spatial data structure(R*tree) to locate points within Eps distance from the core points of the clusters [7]. In addition, another clustering algorithm, GDBSCAN, generalizing the density-based

algorithm DBSCAN is presented in [7], can cluster point instances to both, their numerical and their categorical attributes. PDBSCAN, a parallel version of DBSCAN is presented in [7].

DBCLASD (Distribution Based Clustering of Large Spatial Data sets) eliminates the need for Min Pts and Eps parameters [7]. DBCLASD incrementally augments an initial cluster by its neighboring points as long as the nearest neighbor distance set of the resulting cluster still fits the expected distance distribution. While the distance set of the whole cluster might fit the expected distance distribution, this does not necessarily hold for all subsets of this cluster. Thus, the order of testing the candidates is crucial. OPTICS is introduced in [8], which is an interactive clustering algorithm, works by creating an ordering of the data set representing its density-based clustering structure.

Another density-based algorithm is the DENCLUE [8]. The basic idea of DENCLUE is to model the overall point density analytically as the sum of influence functions of the data points. The algorithm then works by determining the maximum of the overall density function to identify clusters. The algorithm allows a compact mathematical description of arbitrarily shaped clusters in high-dimensional data sets and is significantly faster than the other density based clustering algorithms. Moreover, DENCLUE produces good clustering results even when a large amount of noise is present. In this approach, there are two important parameters, namely σ (influence of a point in its neighborhood) and ξ (significance of density-attractor). Density-attractors are local maxima of the overall density function. FDC algorithm (Fast Density-Based Clustering) is presented in [5] for density-based clustering defined by the density-linked relationship. The clustering in this algorithm is defined by an equivalence relationship on the objects in the database. The complexity of FDC is linear to the size of the database, which is much faster than that of the algorithm DBSCAN[4].

Density based clustering is applicable to spatial data and on more than one attribute to be selected for clustering. Partly, need to enter the number of clusters is solved in few of the algorithms. These are scalable to the database size (mostly).

V. GRID-BASED CLUSTERING

Grid-based clustering algorithms first quantize the clustering space into a finite number of cells and then perform the required operations on the quantized space. Cells that contain more than certain number of points are treated as dense and the dense cells are connected to form the clusters. Some of the grid-based clustering algorithms are: Statistical INformation Grid-based method – STING, Wave Cluster, and CLustering In QUEst – CLIQUE.

STING [8] first divides the spatial area into several levels of rectangular cells in order to form a hierarchical structure. The cells in a high level are composed from the cells in the lower level. It generates a hierarchical structure of the grid cells so as to represent the clustering information at different levels. Although STING generates good clustering results in a short running time, there are two major problems with this algorithm. Firstly, the performance of STING relies on the granularity of the lowest level of the grid structure. Secondly, the resulting clusters are all bounded horizontally or vertically, but never diagonally. This shortcoming might greatly affect the cluster quality.

CLIQUE [8] is another grid-based clustering algorithm that starts by finding all the dense areas in the one-dimensional spaces corresponding to each attribute, and then generates the set of two-dimensional cells that might possibly be dense by looking at dense one-dimensional cells. Generally, CLIQUE generates the possible set of k-dimensional cells that might possibly be dense by looking at dense (k - 1) dimensional cells. CLIQUE produces identical results irrespective of the order in which the input records are presented. In addition, it generates cluster descriptions in the form of DNF expressions for ease of comprehension. Moreover, empirical evaluation shows that CLIQUE scales linearly with the number of instances, and has good scalability as the number of attributes is increased.

Unlike other clustering methods, 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 [8].

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.

VI. MODEL BASED METHODS

Auto Class algorithm uses the Bayesian approach, starting from a random initialization of the parameters, incrementally adjusts them in an attempt to find their maximum likelihood estimates. Moreover, in [9] it is assumed that, in addition to the observed or predictive attributes, there is a hidden variable. This unobserved variable reflects the cluster membership for every case in the data set. Therefore, the data-clustering problem is also an example of supervised learning from incomplete data due to the existence of such a hidden variable. Their approach for learning has been called RBMNs (Recursive Bayesian Multinets).

Another model based method is the SOM net [3]. The SOM net can be thought of as two layers neural network. Each neuron is represented by n-dimensional weight vector, where n is equal to the dimension of the input vectors. The neurons of the SOM are themselves cluster centers map units are used to form bigger clusters iteratively (during training). It is robust, can deal with missing data values and can detect outlier easily from the map, since its distance in the input space from other units is large. The philosophy of the approach is to assume clustering as classification problem with missing data and this algorithm produces arbitrary shaped clusters.

VII. ENSEMBLES OF CLUSTERING ALGORITHMS

Many applications require the clustering of large amounts of high dimensional data. However, most automated clustering techniques do not work effectively and/or efficiently on high dimensional data, i.e. they are likely to miss clusters with certain unexpected characteristics.

Reasons being:

- Difficulty in finding necessary parameters for tuning the clustering algorithms to the specific application.
- It is hard to verify and interpret the high dimensional clusters

A solution is to try to build a solution that is a combination of clustering algorithms (ensembles of clustering algorithms).The theoretical foundation of combining multiple clustering algorithms is still in its early stages. In fact, combining multiple clustering algorithms is a more challenging problem than combining multiple classifiers. In [8] the reason that impede the study of clustering combination has been identified, Cluster ensembles can be formed in a number of different ways , such as using the several clustering techniques of different type, same technique with different configurations, making use of intermediate results of one technique with the other. A split-and-merge strategy is followed in some methods. The first step is to decompose complex data into small, compact clusters using K-means. Data partitions present in these clusters are mapped into a new similarity matrix between patterns, based on a voting mechanism. The idea of combining multiple clustering algorithms of a set of data patterns based on a Weighted Shared nearest neighbors Graph (WSnnG) is introduced in [8]. Due to the increasing size of current databases, constructing efficient distributed clustering algorithms has attracted considerable attention.

Distributed Clustering assume that the objects to be clustered reside on different sites. Instead of transmitting all objects to a central site (also denoted as server) where we can apply standard clustering algorithms to analyze the data, the data are clustered independently on the different local sites (also denoted as clients). In a subsequent step, the central site tries to establish a global clustering based on the local models. Generally, as far as distributed clustering is concerned, there are different scenarios where in the approach work by combining a set of clusters obtained from clustering algorithm,

- Feature-Distributed Clustering (FDC) having partial view of the data features.
- Object-Distributed Clustering (ODC) having access to the whole set of data features and to a limited number of objects.
- Feature/Object-Distributed Clustering (FODC) having access to limited number of objects and/or features of the data.

TABLE I. LIST OF CLUSTERING ALGORITHMS

Algorithm	Data type	Cluster shape	Complexity	DATA SET	Measure	PRIMARY INPUT REQUIRED
K-MEANS	NUMERICAL	SPHERICAL	O(TKN)	LARGE	MEAN	NUMBER OF CLUSTERS
K-MEDOIDS	NUMERICAL	ARBITRARY	O(TKN)	LARGE	MEDOID	NUMBER OF CLUSTERS
CLARA	NUMERICAL	ARBITRARY	O(KS2+K(N-K))	SAMPLE	MEDOID	NUMBER OF CLUSTERS
CLARANS	NUMERICAL	ARBITRARY	O(N2)	SAMPLE	MEDOID	NUMBER OF CLUSTERS
FUZZY K-MEANS	NUMERICAL	ARBITRARY	O(tkN)	LARGE	MEAN	NUMBER OF CLUSTERS
SQUEEZER	MIXED	SPHERICAL	O(kN)		OBJECT SIMILARITY	MINIMUM SIMILARITY
K-PROTOTYPES	MIXED	SPHERICAL	O(tkN)		MODE	NUMBER OF CLUSTERS
FFT	MIXED	SPHERICAL	O(NK)		MEAN	NUMBER OF CLUSTERS
COOLCAT	DISCRETE	SPHERICAL	O(N2)		ENTROPY	NUMBER OF CLUSTERS
FUZZY K-MODES	DISCRETE		O(TKN)		MODES	NUMBER OF CLUSTERS
CLICK	MIXED	LAYERED CLUSTERS	FAST		VERTICES AND EDGES	
BRICH	NUMERICAL	SPHERICAL	O(N)	LARGE	FEATURE TREE	
CURE	NUMERICAL	ARBITRARY	O(N)		SIMILARITY MEASURE	
ROCK	MIXED	GRAPH	O(KN2)	SMALL SIZED	SIMILARITY MEASURE	NUMBER OF CLUSTERS
CHAMELEON	DISCRETE	ARBITRARY	O(N2)		SIMILARITY MEASURE	MIN SIMILARITY
LIMBO	DISCRETE		O(NLOGN)	LARGE	SIMILARITY MEASURE	INFO LOSS
HEMITS			FAST			THRESHOLD
POWER GRAPHS	MIXED		O(ND2)			THRESHOLD
HERDENCE	MIXED		O(N)			
MULIC	MIXED		O(N2)			
DBSCAN	NUMERICAL	ARBITRARY	O(NLOGN)	HIGH DIMENSIONAL	DENSITY BASED	DENSITY THRESHOLD
OPTICS	NUMERICAL	ARBITRARY	O(NLOGN)	HIGH DIMENSIONAL	DENSITY BASED	DENSITY THRESHOLD
DENCLUE	NUMERICAL	ARBITRARY	O(N2)	HIGH DIMENSIONAL	DENSITY BASED	RADIUS
CACTUS	DISCRETE		SCALABLE			NUMBER OF CLUSTERS
STIR	DISCRETE		SCALABLE		WEIGHTS, DISTANCE BASED	NODE WEIGHTS
CLICK	DISCRETE		SCALABLE		OBJECT	
CLOPE	DISCRETE		O(KDN)	HIGH DIMENSIONAL	DISTANCE BASED	
WAVECLUSTER	NUMERICAL	ARBITRARY	O(N)	LOW DIMENSIONAL	WAVE TRANSFORM	WAVELET TRANSFORM
STING	NUMERICAL	RECTANGULAR	O(N)	ANY SIZE	STATISTICAL	STASTICAL
CLIQUE	MIXED	ARBITRARY	O(N)	HIGH DIMENSIONAL	DENSITY BASED	DENSITY THRESHOLD
SOMS	NUMERICAL	ARBITRARY	O(N2)	LOW DIMENSIONAL	OBJECT SIMILARITY	NUMBER OF CLUSTERS
COBWEB	DISCRETE	TREE	O(N2D)	LARGE	ENTROPY	
BILCOM	MIXED	ARBITRARY	O(N2)	LARGE	PROBABILITY	THRESHOLD
AUTOCLASS	MIXED	ARBITRARY	O(KDINT)	0	PROBABILITY	

The table 1 describes about the clustering algorithms and their corresponding characteristics.

VIII. CONCLUSIONS

The main problems in unsupervised learning are guessing the right number of output clusters and stopping criteria like number of iteration to stop specified. These two affect the accuracy and performance of the algorithm. The Robustness in handling noisy data, outliers, Order sensitivity, Dataset size and Shape that affects understanding of the clustering results. The algorithms that can handle numeric data are K-Means, Farthest First Traversal k-center (FFT), K-Medoids, CLARA, CLARANS, Fuzzy k-means, BIRCH, CURE, STING, CLIQUE, DBSCAN, OPTICS, DENCLUE, Wave Cluster , CLIQUE, SOM, etc. The algorithms that can handle discrete

data are K-Modes, Fuzzy k-modes, Squeezer, COOLCAT, ROCK, Chameleon, LIMBO, HIERDENC, MULIC, Projected (subspace) clustering, CACTUS, STIRR, CLICK, CLOPE, COBWEB, etc there is one more category comprising mixed data type which has some methods like K-Prototypes, BILCOM, Auto Class, SVM Clustering, etc. Partition clustering methods are applicable on single attribute provided user enters number of clusters and stopping criteria; they are order sensitive and cannot handle noisy data. They often tend to produce spherical shaped clusters only. Hierarchical clustering is mostly applicable on sample dataset or random sample and sensitive to order of presentation of data. They handle outliers and noisy data better than partition clustering methods and produces arbitrary shaped clusters representing data appropriately in the clusters. Density based clustering is applicable to spatial data and on more than one attribute to be selected for clustering. Partly, need to enter the number of clusters is solved in few of the algorithms. These are scalable to the database size (mostly). 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. The philosophy of the approach is to assume clustering as classification problem with missing data. The clusters produce arbitrary shaped clusters. The ensemble approach make two algorithms work in conjugation. The best approach may be used for achieving best results. It should not be forgotten that no algorithm is a readymade solution for all the issues; unsupervised learning is still waiting for a one shot solution that fits all the bills.

REFERENCES:

[1] HAN, J. and KAMBER, M. 2001. Data Mining. Morgan Kaufmann Publishers

[2] S.Z. Selim, M.A. Ismail, "K Means Type Algorithms: A Generalized Convergence Theorem and characterization of Local Optimality," IEEE Trans Pattern Analysis and Machine Intelligence, pp. 8187, 1984.

[3] Balasubramaniyan R, Huellermeier E, Weskamp N, et al.Clustering of gene expression data using a local shape-based similarity measure. Bioinformatics 2005;21:1069–77.

[4] A. Hinneburg and D.A. Keim, "A General Approach to Clustering in Large Databases with Noise," Knowledge and Information Systems (KAIS), vol. 5, no. 4, pp. 387415, 2003.

[5] Bill Andreopoulos, Aijun An, XiaogangWang and Michael Schroeder ,Advance Access publication February 24, 2009 A roadmap of clustering algorithms: finding a match for a biomedical application in BRIEFINGS IN BIOINFORMATICS. VOL 10. NO 3.

[6] D.T. Pham, S.S. Dimov, C.D. Nguyen, "An Incremental Kmeans Algorithm", Proceedings of the Institution of Mechanical Engineers, Journal of Mechanical Engineering Science, vol. 218, Issue 7, pp.783795, 2004.

[7] Xu R. Survey of clustering algorithms. IEEE Trans. Neural Networks 2005;16.

[8] B¨OHM, C., KAILING, K., KRIEGEL, H.-P., AND KR¨OGER, P. 2004. Density connected clustering with local subspace preferences. In Proceedings of the 4th International Conference on Data Mining (ICDM).

[9] B¨OHM, C., KAILING, K., KR¨OGER, P., AND ZIMEK, A. 2004a. Computing clusters of correlation connected objects. In Proceedings of the ACM International Conference on Management of Data (SIGMOD).