A Survey on Different Clustering Algorithms in Data Mining Technique

P. IndiraPriya, [1] Dr. D.K.Ghosh[2]

[1]*Tagore Engineering College, Chennai, India*
[2]*V.S.B. Engineering College, Karur, India*

ABSTRACT: *Fast retrieval of the relevant information from databases has always been a significant issue. There are many techniques are developed for this purpose; In among data clustering is one of the major technique. The process of creating vital information from a huge amount of data is learning. It can be classified into two such as supervised learning and unsupervised learning. Clustering is a kind of unsupervised data mining technique. It describes the general working behavior, the methodologies followed by these approaches and the parameters which affect the performance of these algorithms. In classifying web pages, the similarity between web pages is a very important feature. The main objective of this paper is to gather more core concepts and techniques in the large subset of cluster analysis.*

Keywords: *Clustering, Unsupervised Learning Web Pages, Classifications.*

I. Introduction

Now a day, people come across a huge amount of information and store or represent it as data. One of the vital means in dealing with these data is to classify or group them into a set of categories or clusters. Clustering involves creating groups of objects that are similar, and those that are dissimilar. The clustering problem lies in finding groups of similar objects in the data. The similarity between the objects is measured with the use of a similarity function. Clustering is especially useful for organizing documents, to improve retrieval and support browsing. Clustering is often confused with classification, but there is some difference between the two. In classification, the objects are assigned to pre defined classes, whereas in clustering the classes are also to be defined. To be Precise, Data Clustering is a technique in which, the information that is logically similar is physically stored together. In order to increase the efficiency in the database system the numbers of disk accesses are to be minimized. In clustering, objects having similar properties are placed in one class, and a single access to the disk makes the entire class available. Clustering algorithms can be applied in many areas, for instance, marketing, biology, libraries, insurance, city-planning, earthquakes, and www document classification.

1.1 Clustering

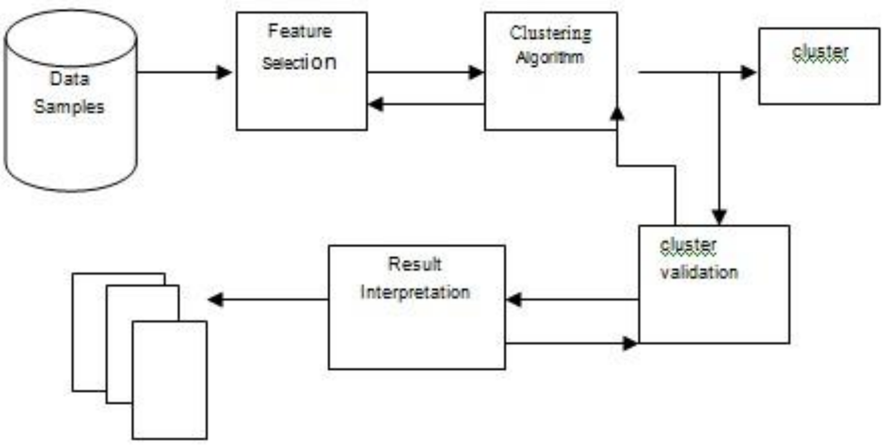
Clustering can be considered the most important unsupervised learning problem; so, as with every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. The process of organizing objects into groups whose members are similar in some way is a cluster. A collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. Two important topics are: (1) Different ways to group a set of objects into a set cluster (2) types of Clusters.

The analysis of cluster is to identify and classifies objects, individuals or variables, on the basis of the similarities. It seeks to minimize within-group variance and maximize between-group variance. The result of the cluster analysis is a number of heterogeneous groups with homogeneous contents. The individuals within a single group are similar rather than substantial differences between the groups.

The Cluster analysis groups data objects based only on the information found in the data that describes the objects and their relationships. The goal is that the objects within a group be similar (or related) to one another and different from (or unrelated to) the objects in the other groups. The greater similarity (or homogeneity) of clustering is within a group, and the greater the difference between groups, the better or more distinct the clustering [8]. Data may be thought of as points in a space where the axes correspond to the variables. The cluster analysis divides the space into regions, characteristic of the groups found in the data. The main advantage of a clustered solution is automatic recovery from failure, that is, recovery without user intervention. The disadvantages of clustering are complexity and inability to recover from database corruption.
An ordered list of objects, which have some common characteristics of cluster. The objects belong to an interval [a, b], in our case [0, 1] [2]. The distance between two clusters involves some or all elements of the two clusters. The clustering method determines how the distance should be computed [1]. A similarity measure SIMILAR (D_i, D_j) can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms, and 1 when perfect agreement is detected. Intermediate values are obtained for cases of partial agreement [1]. If the similarity measure is computed for all pairs of documents (D_i, D_j) except when i=j, an average value AVERAGE SIMILARITY is obtainable. Specifically, AVERAGE SIMILARITY = CONSTANT SIMILAR (D_i, D_j), where i=1,2,….n and j=1,2,….n and i< > j. The lowest possible input value of similarity is required to join two objects in one cluster. The similarity between objects calculated by the function SIMILAR ($D_{i,}$,D_j), represented in the form of a matrix is called a similarity matrix. The dissimilarity coefficient of two clusters is defined as the distance between them. The smaller the value of the dissimilarity coefficient, the more similar the two clusters are.

The first document or object of a cluster is defined as the initiator of that cluster, i.e., similarity of every incoming object's is compared with the initiator. The initiator is called the cluster seed. The procedure of the cluster analysis with four basic steps is as follows:

1) ***Feature selection or extraction.*** As pointed out by Jain *et al.* [5], [6] and Bishop [7], feature selection chooses distinguishing features from a set of candidates, while feature extraction utilizes some transformations to generate useful and novel features from the original ones. An elegant selection of features can greatly decrease the workload, and simplify the subsequent design process. Generally, ideal features should be of use in distinguishing patterns belonging to different clusters, immune to noise, easy to extract and interpret. An elaborate discussion on feature extraction, in the context of data visualization and dimensionality reduction. More information on feature selection can be found in [7], [5], and [9].



Knowledge

2) ***Clustering algorithm design or selection.*** This step is usually combined with the selection of a corresponding proximity measure, and the construction of a criterion function. Patterns are grouped according to whether they resemble one another. Obviously, the proximity measure directly affects the formation of the resulting clusters. Almost all clustering algorithms are explicitly or implicitly connected to some definition of the proximity measure. Some algorithms even work directly on the proximity matrix. Once a proximity measure is chosen, the construction of a clustering criterion function makes the partition of clusters an optimization problem, which is well defined mathematically, and has rich solutions in the literature. Clustering is ubiquitous, and a wealth of clustering algorithms has been developed to solve different problems in specific fields. It has been very difficult to develop a unified framework for reasoning about it (clustering) at a technical level, and profoundly diverse approaches to clustering [10], as proved through an impossibility theorem. Therefore, it is important to carefully investigate the characteristics of the problem on hand, in order to select or design an appropriate clustering strategy.

3) ***Cluster validation.*** Given a data set, each clustering algorithm can always generate a division, no matter whether the structure exists or not. Moreover, different approaches usually lead to different clusters; and even for the same algorithm, parameter identification or the presentation order of the input patterns may affect the final results. Therefore, effective evaluation standards and criteria are important to provide the users with a degree of confidence, for the clustering results derived from the used algorithms. These assessments should be objective and have no preferences to any algorithm. Also, they should be useful for answering questions like how many clusters are hidden in the data, whether the clusters obtained are meaningful or just artifacts of the algorithms. Generally, there are three categories of testing criteria: external indices, internal indices, and relative indices. These are defined on three types of clustering structures, known as partitional clustering, hierarchical clustering, and individual clusters [11]. Tests for a situation, where no clustering structure exists in the data, are also considered [12], but seldom used, since users are confident of the presence of clusters. External indices are based on some pre specified structure, which is a reflection of prior information on the data, and used as a standard to validate the clustering solutions. Internal tests are not dependent on external information (prior knowledge). On the contrary, they examine the clustering structure directly from the original data. Relative criteria place the emphasis on the comparison of different clustering structures, in order to provide a reference, to decide which one may best reveal the characteristics of the objects. Shall not survey the topic in depth, but refer interested readers to [13], [14]. Approaches to fuzzy clustering validity are reported in [16], [17], [18].

4) ***Results interpretation.*** The ultimate goal of clustering is to provide users with meaningful insights into the original data, so that they can effectively solve the problems encountered. Experts in the relevant fields interpret the data partition. Further analyzes, even experiments, may be required to guarantee the reliability of the extracted knowledge.

1.2 Classification

Classification plays a vital role in many information management and retrieval tasks. On the Web, the classification of page content is essential to focused crawling, to the assisted development of web directories, to topic-specific web link analysis, and to the analysis of the topical structure of the Web. Web page classification can also help improve the quality of web search. Web page classification, also known as web page categorization, is the process of assigning a web page to one or more predefined category labels. Classification is often posed as a supervised learning problem (Mitchell 1997) in which a set of labeled data is used to train a classifier, which can be applied to label future examples.

The general problem of web page classification can be divided into multiple sub-problems: subject, functional, sentiment, and other types of classification. Subject classification concerns the subject or topic of a Web page. For example, judging whether a page is about "arts", "business" or "sports" is an instance of subject classification. Functional classification cares about the role that the Web page plays. For example, deciding a page to be a "personal homepage", "course page" or "admission page" is an instance of functional classification. Sentiment classification focuses on the opinion that is presented in a web page, i.e., the author's attitude about some particular topic. Other types of classification include genre classification (e.g., (zu Eissen and Stein 2004)), search engine spam classification (e.g., (Gyongyi and Garcia-Molina 2005b; Castillo, Donato, Gionis, Murdock, and Silvestri 2007)) and so on.

Based on the number of classes in the problem, classification can be divided into binary classification and multi-class classification, where binary classification categorizes instances into exactly one of two classes; multi-class classification deals with more than two classes. Based on the number of classes that can be assigned to an instance, classification can be divided into single-label classification and multi-label classification. In single-label classification, one and only one class label is to be assigned to each instance, while in multi-label classification; more than one class can be assigned to an instance. If a problem is multi-class, say four-class classification, it means four classes are involved, say Arts, Business, Computers, and Sports. It can be either single-label, where exactly one class label can be assigned to an instance, or multi-label, where an instance can belong to any one, two, or all of the classes. Based on the type of class assignment, classification can be divided into hard classification and soft classification. In hard classification, an instance can either be or not be in a particular class, without an intermediate state; while in soft classification, an instance can be predicted to be in some class with some likelihood (often a probability distribution across all classes).

Based on the organization of categories, web page classification can also be divided into flat classification and hierarchical classification. In flat classification, categories are considered parallel, i.e., one category does not supersede another, while in hierarchical classification, the categories are organized in a hierarchical tree-like structure, in which each category may have a number of subcategories.

Clustering can be in the form of classification, in that it creates a labeling of objects with class (cluster) labels. Classification means a supervised classification; i.e., new, unlabeled objects are assigned a class label using developed objects with known class labels. The term segmentation and partitioning are sometimes used as synonyms for clustering. The partitioning is often used in connection with techniques that divide graphs into sub graphs and that are not strongly connected to clustering. Segmentation often refers to the division of data into group using simple techniques; eg., an image can be split into segments based only on pixel intensity and color, or people can be divided into groups based on their income. Clustering is a type of classification imposed on a finite set of objects. The relationship between objects is represented in a proximity matrix, in which rows and columns correspond to objects. The proximity matrix is the one and only input to a clustering algorithm.

In this paper, various clustering algorithms in data mining are discussed. A new approach for to improve the prediction accuracy of the clustering algorithms is proposed.

II. Approaches

2.1 Types of Clustering Algorithms

Clustering is a division of data into groups of similar objects [3]. The clustering algorithm can be divided into five categories, viz, Hierarchical, Partition, Spectral, Grid based and Density based clustering algorithms.

2.1.1 Hierarchical Clustering Algorithm

The hierarchical clustering algorithm is a group of data objects forming a tree shaped structure. It can be broadly classified into agglomerative hierarchical clustering and divisive hierarchical clustering. In the agglomerative approach, which is also called as the bottom up approach, each data point is considered to be a separate cluster, and on each iteration the clusters are merged, based on a criterion. The merging can be done by using the single link, complete link, centroid or wards method. In the divisive approach all data points are considered as a single cluster, and they are split into a number of clusters, based on certain criteria, and this is called as the top down approach. Examples of this algorithms are LEGCLUST [22], BRICH [19] (Balance Iterative Reducing and Clustering using Hierarchies), CURE (Cluster Using REpresentatives) [20], and Chemeleon [1].

The construction of a hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the 2 closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains, return to step 2.

Individual methods are characterized by the definition used for the identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged. There are some general approaches to the implementation of this algorithm; these being stored matrix and stored data, are discussed below:

- In the second matrix approach , an N*N matrix containing all pair wise distance values is first created, and updated as new clusters are formed. This approach has at least an O(n*n) time requirement, rising to O(n^3) if a simple serial scan of dissimilarity matrix is used to identify the points, which need to be fused in each agglomeration, a serious limitation for large N.
- The stored data approach requires the recalculation of the pair wise dissimilarity values for each of the N-1 agglomerations, and the O (N) space requirement is therefore achieved at the expense of an O (N^3) time requirement.

The advantages of hierarchical clustering include embedded flexibility regarding the level of granularity, and Ease of handling of any forms of similarity or distance. Consequently, its applicability to any attributes types and its logical structure, make it easy to read and interpret. The disadvantages of hierarchical clustering are related to the vagueness of termination criteria, the fact that most hierarchical algorithms do not revisit once constructed, (intermediate) clusters with the purpose of their improvement; and that they are relatively unstable and unreliable, i.e., the first combination or separation of objects, which may be based on a small difference in the criterion, will constrain the rest of the analysis.

2.1.2 Spectral Clustering Algorithm

Spectral clustering refers to a class of techniques, which relies on the Eigen structure of a similarity matrix. Clusters are formed by partitioning data points using the similarity matrix. Any spectral clustering algorithm will have three main stages [23]. They are preprocessing, spectral mapping and post mapping. Preprocessing deals with the construction of the similarity matrix. Spectral Mapping deals with the construction of Eigen vectors for the similarity matrix. Post Processing deals with the grouping of data points.

The advantages of the spectral clustering algorithm are: strong assumptions on the cluster shape are not made; it is simple to implement and objective; it does not consider local optima; it is statistically consistent and works faster. The major drawback of this approach is that it exhibits high computational complexity. For large data set it requires O (n3), where n is the number of data points [17]. Examples of this algorithm are, SM(Shi and Malik) algorithm, KVV (Kannan,VempalaandVetta) algorithm, and NJW (Ng, Jordan and Weiss)algorithm [22].

2.1.3 Grid based Clustering Algorithm

The grid based algorithm quant sizes the object space into a finite number of cells, that forms a grid structure [1].Operations are done on these grids. The advantage of this method is its lower processing time. Clustering complexity is based on the number of populated grid cells, and does not depend on the number of objects in the dataset. The major features of this algorithm are, no distance computations, Clustering is performed on summarized data points, Shapes are limited to the union of grid-cells, and the complexity of the algorithm is usually O(Number of populated grid-cells). STING [1] is an example of this algorithm.

2.1.4 Density based Clustering Algorithm

The density based algorithm allows the given cluster to continue to grow as long as the density in the neighbor hood exceeds a certain threshold [4]. This algorithm is suitable for handling noise in the dataset. The following points are enumerated as the features of this algorithm: it handles clusters of arbitrary shape, Handles noise, needs only one scan of the input dataset, and the density parameters to be initialized. DBSCAN, DENCLUE and OPTICS [4] are examples of this algorithm.

Density-Based Connectivity

The crucial concepts of this section are density and connectivity, both measured in terms of local distribution of nearest neighbors.

The algorithm DBSCAN (Density Based Spatial Clustering of Applications with Noise) targeting low-dimensional spatial data is the major representative of this category.

Two input parameters and Min Pts are used to define:
1) An -neighborhood ( ) { | (,) } N x = y_ X d x y _ of the point x,
2) A core object (a point with a neighborhood consisting of more than Min Pts points)
3) A concept of a point y density-reachable from a core object x (a finite sequence of core objects between x and y exists such that each next belongs to an - neighborhood of its predecessor)
4) A density-connectivity of two points x, y (they should be density-reachable from a common core object).

Density Functions

Hinneburg & Keim [1998] shifted the emphasis from computing densities pinned to data points to computing density functions defined over the underlying attribute space. They proposed the algorithm DENCLUE (DENsity-based CLUstEring). Along with DBCLASD, it has a firm mathematical foundation. DENCLUE uses a density function.

$$f^{D}(x) = {}_{y\text{€}\,D}\, f(x, y) \qquad (1)$$

That is the superposition of several influence functions. When the f-term depends on x y, the formula can be recognized as a convolution with a kernel. Examples include a square wave function $f(x, y) = \Theta(\|x - y\|\ \sigma)$ equal to 1, if the distance between x and y is less than or equal to 0, and a Gaussian influence function $f(x,y) = e^{-\|x\text{-}xy\|2\ /2\ \sigma2}$.This provides a high level of generality: the first example leads to DBSCAN, the second one to k-means clusters! Both examples depend on parameter σ. Restricting the summation to $D = \{y :|\ |\ x - y\ \|< k\ \sigma\ \}c\ X$ enables a practical implementation. DENCLUE concentrates on the local maxima of the density functions called density-attractors, and uses the flavor of the gradient hill-climbing technique for finding them. In addition to center-defined clusters, arbitrary-shape clusters are defined as continuations along sequences of points whose local densities are no less than the prescribed threshold £. The algorithm is stable with respect to outliers and authors show how to choose parameters £ and σ. DENCLUE scales well, since at its initial stage it builds a map of hyper-rectangle cubes with an edge length 2 σ. For this reason, the algorithm can be classified as a grid-based method.

The advantages of this density function in clustering are, that it does not require a-priori specification of the number of clusters, is able to identify noise data while clustering, and the DBSCAN algorithm is able to find arbitrarily sized and arbitrarily shaped clusters. The disadvantages of this density function in clustering are that the DBSCAN algorithm fails in the case of varying density clusters and in the case of a neck type of dataset.

2.1.5 Partition Clustering Algorithm

Partitioning methods generally result in a set of M clusters, each object belonging to one cluster. Each cluster may be represented by a centroid or a cluster representative; this is some sort of a summary description of all the objects contained in a cluster. The precise form of this description will depend on the type of the object which is being clustered. In cases where real-valued data is available, the arithmetic mean of the attribute vectors for all objects within a cluster provides an appropriate representative; alternative types of centroid may be required in other cases; e.g., a cluster of documents can be represented by a list of those keywords that occur in some minimum number of documents within a cluster. If the number of clusters is large, the centroids can be further clustered to produce a hierarchy within a dataset.

The partition clustering algorithm splits the data points into k partition, where each partition represents a cluster. The partition is done based on certain objective functions. One such criterion function is minimizing the square error criterion which is computed as,

$$E = \Sigma\,\Sigma\ ||\ p - mi\ ||\ 2 \qquad (2)$$

Where p is the point in a cluster and mi is the mean of the cluster. The cluster should exhibit two properties; they are (1) each group must contain at least one object (2) each object must belong to exactly one group. The main drawback of this algorithm [3] is whenever a point is close to the center of another cluster; it gives a poor result due to overlapping of the data points.

Single Pass is a very simple partition method; it creates a partitioned dataset in three steps. First, it makes the first object the centroid for the first cluster. For the next object, it calculates the similarity, S, with each existing cluster centroid, using some similarity coefficient. Finally, if the highest calculated S is greater than some specified threshold value, it adds the object to the corresponding cluster, and re determines the centroid; otherwise, it uses the object to initiate a new cluster. If any objects remain to be clustered, it returns to step 2.

As its name implies, this method requires only one pass through the dataset; the time requirements are typically of the order *O(NlogN)* for order *O(logN)* clusters. This makes it a very efficient clustering method for a serial processor. A disadvantage is that the resulting clusters are not independent of the order in which the documents are processed, with the first clusters formed usually being larger than those created later in the clustering run.

K- Means Clustering Algorithm is one of the partition based clustering algorithms. The advantages of the simple K means algorithm. That it is easy to implement and works with any of the standard norms. It allows straight forward parallelization; and it is insensitive with respect to data ordering. The disadvantages of the K means algorithm are as follows. The results strongly depend on the initial guess of the centroids. The local optimum (computed for a cluster) does not need to be a global optimum (overall clustering of a data set). It is not obvious what the good number K is in each case, and the process is, with respect to the outlier.

2.2 Soft Clustering

2.2.1 Fuzzy K Means Clustering

Fuzzy clustering allows each feature vector to belong to more than one cluster with different membership degrees (between 0 and 1), and vague or fuzzy boundaries between clusters. Fuzzy clustering is often used in modeling (fuzzy modeling, neural networks, rule-based systems), where the clusters are sought as a structural setup for the ensuing modelling activities. In this case, the viewpoints can be formed as some points that are located at the boundaries of the range of input variables, so that we achieve a comprehensive "coverage" of the input space, and this way, the models "spanned" over these information granules can be highly representative. The clusters exhibit a certain "crowding" tendency. It is very unlikely to see the clusters positioned at the extreme values of the input variables and, thus, represent these regions when it comes to the construction of the model. To elaborate on this effect in more detail and high light its implications to system modeling, let us consider a rule-based model that is governed by the rules of the form if **x** is A_i . . . then y is B_i and if **x** is A_c . . . then y is B_c , where *Ai* and *Bi* are fuzzy sets, that are defined in the input and output spaces, respectively. Quite commonly, these fuzzy sets are developed through fuzzy clustering, and the centers of the clusters (prototypes) are the modes of the fuzzy sets A_i and B_i. Alluding to the prototypes formed through clustering, we note that they are formed in the aggregate input–output space, i.e., [v_i m_i]. Three difficulties exist in fuzzy clustering. First, the optimal number of clusters K to be created, has to be determined (the number of clusters cannot always be defined *a* priori and a good cluster validity criterion has to be found). Second, the character and location of the cluster prototypes (centers) is not necessarily known *a priori*, and initial guesses have to be made. Third, the data characterized by large variability in the cluster shape, cluster density, and the number of points (feature vectors) in different clusters, have to be handled.

2.2.2 Fuzzy C Means Clustering

This algorithm works by assigning the membership to each data point corresponding to each cluster center, on the basis of the distance between the cluster center and the data point. The nearer data is to the cluster center, the more is its membership towards the particular cluster center. Clearly, the summation of the membership of each data point should be equal to one. The advantages of this clustering algorithm are it gives the best result for an overlapped data set, and a comparatively better then k-means algorithm. Unlike the k-means, where the data point must exclusively belong to one cluster center, here the data point is assigned a membership to each cluster center, as a result of which the data point may belong to more than one cluster center. The disadvantages of the clustering algorithm are, Apriori specification of the number of clusters; with a lower value of β get a better result but at the expense of more number of iterations and the Euclidean distance measures can unequally weight underlying factors.

2.2.3 New Weighted Fuzzy C Means Clustering

A new weighted fuzzy C-means (NW-FCM) algorithm was proposed by Chih-Cheng Hung et al [26]; it is used to improve the performance of both the FCM models for high dimensional multiclass pattern recognition problems. The methodology used in NW-FCM is the concept of the weighted mean from the non parametric weighted feature extraction (NWFE) and the cluster mean from the discriminate analysis feature extraction (DAFE).

2.3 Neural Network Based Clustering

Neural networks-based clustering has been dominated by SOFMs, and the adaptive resonance theory (ART), both of which are reviewed here. The objective of SOFM is to represent high-dimensional input patterns with prototype vectors that can be visualized in a usually two-dimensional lattice structure [21]. Each unit in the lattice is called a neuron, and the adjacent neurons are connected to each other, which gives a clear topology of how the network fits itself in to the input space. The input patterns are fully connected to all neurons via adaptable weights, and during the training process, the neighboring input patterns are projected into the lattice, corresponding to the adjacent neurons. In this sense, some authors prefer to think of SOFM as a method to display the latent data structure in a visual way rather than a clustering approach [24].

The merits of neural network based clustering are, the input space density approximation and independence of the order of input patterns and SOFM need to predefine the size of the lattice, i.e., the number of clusters, which is unknown in most circumstances. The de Merits of this neural network based clustering is, it may suffer from input space density misrepresentation [25], where areas of low pattern density may be over-represented, and areas of high density under-represented.

2.4 Genetic Based Clustering Algorithms

2.4.1 Genetic K-Means Algorithm

K. Krishna and M. Narasimha Murty proposed a novel hybrid genetic algorithm (GA) that finds a globally optimal partition of a given data into a specified number of clusters. GAs used earlier in clustering, employ either an expensive crossover operator to generate valid child chromosomes from parent chromosomes, or a costly fitness function or both. To circumvent these expensive operations, they hybridized the GA with a classical gradient descent algorithm used in clustering, viz., the K-means algorithm. Hence, the name genetic K-means algorithm (GKA). They defined the K-means operator, one-step of the K-means algorithm, and used it in GKA as a search operator, instead of crossover. They also defined a biased mutation operator specific to clustering, called distance-based-mutation. Using the finite Markov chain theory, they proved that the GKA converges to a global optimum. It is observed in the simulations that the GKA converges to the best known optimum, corresponding to the given data, in concurrence with the convergence result. It is also observed that the GKA searches faster than some of the other evolutionary algorithms used for clustering. The advantage of the genetic k means clustering algorithm is that it is faster than some of the other clustering algorithms.

2.4.2 Immune Genetic Algorithm based Fuzzy K-means Clustering Algorithm

Chengjie Gu et al [27] proposed a Fuzzy Kernel K Means clustering method based on the immune Genetic algorithm (IGA-FKKM). The Dependence of the fuzzy k means clustering on the distribution of sample was eliminated, with the introduction of the kernel function in this approach. The immune genetic algorithm has been used to suppress fluctuations that occurred at later evolvement and to avoid the local optimum. This algorithm provides the global optimum and higher cluster accuracy.

2.4.3 Immune Genetic Algorithm based Novel Weighted Fuzzy C-means Clustering Algorithm

A Novel Weighted Fuzzy C-Means clustering method, based on the Immune Genetic Algorithm (IGA-NWFCM) was proposed by S.Ganapthy et al [28] for effective intrusion detection. Hence, it improves the performance of the existing techniques to solve the high dimensional multiclass problems. Moreover, the probability of obtaining the global value is increased by the application of the immune genetic algorithm. It provides high classification accuracy, stability, and probability of gaining the global optimum value.

III. Comparative Analysis

The computational complexity of some typical and classical clustering algorithms in Table 1 with several newly proposed approaches specifically designed to deal with large-scale data sets.

Table 1 Computational Complexity of Clustering Algorithms

Clustering Algorithm	Complexity	Capability of tackling high dimensional data
K – means	O(NKd) time O(N+ K) (space)	No
Fuzzy c means	Near O(N)	No
Hierarchical clustering	$O(N^2)$ (time) $O(N^2)$ (space)	No
CLARA	$O(K(40+K)^2+K(N-K))^+$ (time)	No
CLARANS	Quadratic in total performance	No
BIRCH	O(N) (time)	No
DBSCAN	O(N log N) (time)	No
CURE	$O(N^2_{sample}logN_{sample})$ (time) $O(N_{sample})$ (space)	Yes

IV. Proposed Approach

The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how does one decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion, which would be independent of the final aim of the clustering. Consequently, it is the user who must supply this criterion, in such a way that the result of the clustering will suit this needs. For instance, it could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters”, and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

In this review, the clustering scalability and efficiency of various clustering algorithm have been analyzed. This system propose a different Clustering algorithms for data sets appearing in statistics, computer science, and machine learning, and illustrate their applications in some benchmark data sets, the traveling salesman problem, and bioinformatics, a new field attracting intensive efforts. The results of different clustering depict the efficiency of the method. Because of the computation overhead in constructing dissimilarity matrix. There is also some scope for applying the clustering procedure to large datasets. In large datasets, the clustering efficiency is degraded and also need to improve time and scalability values. So to probe novel approaches for making efficient clustering schemas.

V. Conclusion

The cluster analysis examines unlabeled data, by either constructing a hierarchical structure, or forming a set of groups, according to a pre specified number. In this paper, an attempt has been made to give the basic concept of clustering, by first providing the definition of different clustering algorithms and some related terms. The soft clustering technique and hierarchical method of clustering were explained. The main focus was on these clustering algorithms, and a review of a wide variety of approaches that are mentioned in the literature. These algorithms evolve from different research communities, and these methods reveal that each of them has advantages and disadvantages. The drawback of the k-means algorithm is to find the optimal k value and initial centroid for each cluster. This is overcome by applying concepts, such as fuzzy algorithm.

References

[1], Jiawei Han, MichelineKamber, “Data Mining Concepts and Techniques” Elsevier Publication.

[2], AthmanBouguettaya "On Line Clustering", IEEE Transaction on Knowledge and Data Engineering Volume 8, No. 2, April 1996.

[3], P. Berkhin, 2002. Survey of Clustering Data Mining Techniques. Ttechnical report, AccrueSoftware, San Jose, Cailf.

[4], Jiawei Han, Micheline Kamber, “Data Mining Concepts and Techniques” Elsevier Publication.

[5], A. Jain, R. Duin, and J. Mao, “Statistical pattern recognition: A review,” IEEE Trans. Pattern Anal. Mach.Intell., vol. 22, no. 1, pp. 4–37, 2000.

[6], A. Jain, M. Murty, and P. Flynn, “Data clustering: A review,” ACM Comput. Surv., vol. 31, no. 3, pp. 264–323, 1999

[7], Bishop, Neural Networks for Pattern Recognition. New York: Oxford Univ. Press, 1995.

[8], Bruce Moxon "Defining Data Mining, The Hows and Whys of Data Mining, and How It Differs From Other Analytical Techniques" Online Addition of DBMS Data Warehouse Supplement, August 1996.

[9], Handbook of Pattern Recognition and Computer Vision, C. Chen, L. Pau, and P. Wang, Eds., World Scientific, Singapore, 1993, pp. 61–124. J. Sklansky and W. Siedlecki, “Large-scale feature selection”.

[10], J. Kleinberg, “An impossibility theorem for clustering,” in Proc. 2002 Conf. Advances in Neural Information Processing Systems, vol. 15, 2002, pp. 463–470.

[11], A. Jain and R. Dubes, Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[12], A. Gordon, “Cluster validation,” in Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Bada, Eds. New York: Springer-Verlag, 1998, pp. 22–39.

[13], Handbook of Pattern Recognition and Computer Vision, C. Chen, L. Pau, and P.Wang, Eds.,World Scientific, Singapore, 1993, pp. 3–32. R. Dubes, “Cluster analysis and related issue”.

[14], A. Gordon, “Cluster validation,” in Data Science, Classification, and Related Methods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H. Bock, and Y. Bada, Eds. New York: Springer-Verlag, 1998, pp. 22–39.

[15], S. Bandyopadhyay and U. Maulik, “Nonparametric genetic clustering: Comparison of validity indices,” IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 31, no. 1, pp. 120–125, Feb. 2001.

[16], R. Davé and R. Krishnapuram, “Robust clustering methods: A unified view,” IEEE Trans. Fuzzy Syst., vol. 5, no. 2, pp. 270–293, May 1997.

[17], A. Geva, “Hierarchical unsupervised fuzzy clustering,” IEEE Trans. Fuzzy Syst., vol. 7, no. 6, pp. 723–733, Dec. 1999.

[18], R. Hammah and J. Curran, “Validity measures for the fuzzy cluster analysis of orientations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 12, pp. 1467–1472, Dec. 2000.

[19], M. Livny, R.Ramakrishnan, T. Zhang, 1996. BIRCH: An Efficient Clustering Method for VeryLarge Databases. Proceeding ACMSIGMOD Workshop on Research Issues on Data Mining andKnowledge Discovery: 103-114.

[20], S. Guha, R. Rastogi, and K. Shim, 1998. CURE: An Efficient Clustering Algorithm for Large Databases. Proc. ACM Int’l Conf. Management of Data: 73-84.

[21], T. Kohonen, “The self-organizing map,” Proc. IEEE, vol. 78, no. 9, pp.1464–1480, Sep. 1990.

[22], Santos, J.M, de SA, J.M, Alexandre, L.A, 2008. LEGClust- A Clustering Algorithm based on Layered Entropic subgraph. Pattern Analysis and Machine Intelligence, IEEE Transactions: 62-75.

[23], M Meila, D Verma, 2001. Comparison of spectral clustering algorithm. University of Washington, Technical report

[24], N. Pal, J. Bezdek, and E. Tsao, “Generalized clustering networks andKohonen’s self-organizing scheme,” IEEE Trans. Neural Netw., vol. 4, no. 4, pp. 549–557, Jul. 1993.

[25], S. Haykin, Neural Networks: A Comprehensive Foundation, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1999.

[26], Chih-Cheng Hung, Sameer Kulkarni, Bor-Chen Kuo, “A New Weighted Fuzzy C-Means Clustering Algorithm for Remotely Sensed Image Classification”, IEEE Journal of Selected Topics in Signal Processing, Vol. 5, No.3, pp. 543-553, 2011.

[27], Chengjie GU, Shunyi ZHANG, Kai LIU, He Huang, “Fuzzy Kernal K-Means Clustering Method Based on Immune Genetic Algorithm”, Journal of Computational Information Systems, Vol. 7, No. 1, pp. 221-231, 2011.

[28], S.Ganapathy, K.Kulothungan, P.Yogesh, A.Kannan, “ A Novel Weighted Fuzzy C-Means Clustering Based on Immune Genetic Algorithm for Intrusion Detection”, Proceeding Engineering, Elsevier, Vol. 38, pp. 1750-1757, 2012.