A Survey on Data Clustering

Garima Singhal and Sahadev Roy

Abstract

Data Clustering is a method used to group the unlabelled data based on its similarity according to their type, specification and other properties. The current agglomeration focuses on those approaches which help to retrieve and categorize the data based on processing speed, size of data it can support, complexity and memory requirement.

Navigation through this huge unlabelled collection of data presents a challenge for researchers to select an optimal clustering technique. This paper presents a survey report based on analytical responses obtained from existing data clustering algorithms in order to ease the search and to help the audience to select appropriate clustering algorithm according to their requirement. The algorithms which are covered in this paper have application in pattern recognition, image processing, data mining, machine learning and Artificial intelligence. This survey is also useful for those readers who view it as an accessible introduction to the mature content on computer advancements and its development.

Keywords
Artificial Neural Networks (ANN);
Fuzzy c-means algorithm (FCM);
Genetic algorithm (GA);
Self-Organising map (SOM);
Tabu Search (TS).

I. INTRODUCTION

Data Analysis requires extensive research to agglomerate the data and use it in computing applications while in design phase or in on-line operation. Based on the appropriate data source models, data analysis can be dichotomized as either exploratory or confirmatory. But in both the approaches key element is grouping of data based on goodness of fit according to the postulated model. The patterns are usually represented as a vector of measurements or a point in multi-dimensional space. After clustering, patterns belonging to same cluster are more similar to each other than the patterns in adjacent clusters.

Clustering is basically an unsupervised classification of data while supervised classification comes under discriminant analysis. In supervised classification patterns are a group of labeled data and new patterns has to be grouped according to available classes. In unsupervised clustering, patterns are a collection of unlabeled data and they have to be labeled to make more meaningful clusters [1]. These labels are not fixed but are data driven i.e. they solely depends upon data and grouped themselves dynamically according to incoming data.

The process of clustering can be viewed as a five step process [2]:
1. Pattern representation or feature extraction: refers to number of patterns or classes available to cluster algorithm and identifying most appropriate datasets to use in clustering.
2. Definition of pattern proximity in accordance to the defined data domain: It is usually measured by distance function. It calculates the vector distance to define similarity in various points [3].
3. Clustering or grouping: refers to the way clustering classes or methods like hierarchal, Partitional, or fuzzy logic, probabilistic etc.
4. Data Abstraction: Extracting data into more meaningful pattern clusters. It could be a simple or compact representation.
5. Output Assessment: Clustering algorithm's performance is observed at this step. An Algorithm is said to be valid if it doesn't produce clusters by chance or in a default manner.

A single clustering technique cannot be used with all types of data. Each new clustering algorithm performs slightly better than the existing one on some specific distribution of patterns. Other terms which are inter-changeably used with clustering are unsupervised learning, numerical taxonomy, vector quantization, learning by observation and spatial analysis.

The outline of this paper is organized as follow: Detailed taxonomy and key parameters used in comparison are described in section II. Then individual clustering methodologies are briefed in section III. In section IV, a comparative analysis is obtained based on its application and limitation. Section V addresses some real time issues occurs with large data sets and finally section VI concludes the paper.

II. CLUSTERING TAXONOMY

Basically clustering algorithm can be classified in seven major categories. The detailed taxonomy of clustering strategy is presented as below:

1. Hierarchical Clustering algorithm
a) Single link
b) Complete link and
c) Minimum variance

2. Partitional Algorithm
   A) Squared Error Algorithm
   B) Graph-theoretic Clustering

3. Nearest Neighbor clustering
4. Fuzzy clustering
5. Representation of clusters
6. Artificial Neural Networks
7. Evolutionary Approaches

Apart from the performance, working and specification analysis; following parameters must also be considered to inspect a clustering algorithm regardless of its position in clustering taxonomy.

1. Agglomerative Vs Divisive: It refers to the structure of algorithm. Agglomeration means staring from separate clusters for each pattern and successively merging them according to their properties. Divisive means staring from a single cluster containing all patterns and then splitting them successively into smaller clusters.

2. Monolithic Vs Polythetic: It refers to the sequential or simultaneous use of features to perform clustering on the data. In monolithic each feature is considered sequentially and applied on the data while in Polythetic, all features are considered together and thus it performs faster for general data sets. Monolithic is slow to use but has high performance with very large data sets.

3. Hard Vs Fuzzy: It refers to the way algorithm operates and presents its output. Hard clustering represent each cluster as separate entity but in Fuzzy, clusters share some degree of membership with other clusters based on their distance vectors and hence are not completely separated from each other.

4. Deterministic Vs Stochastic: This refers to the way algorithm search its state space to label the clusters. The search can be random or deterministic in nature.

5. Incremental Vs non incremental: This issue comes into effect with very large data sets. When there is a constraint on time and space complexity of the algorithm then minimization of data structures and number of patterns scanned or executed at a time has to be considered.

### III. Clustering Methodology

The major clustering algorithms are briefed in the following section.

**A. Hierarchical Clustering Algorithm**

These clustering algorithms yield the output in a dendrogram form i.e. showing the similarity between patterns in a nested form. There are three variants of hierarchical clustering in which the process of assigning clusters depends upon the underlying concept the algorithm is based on. These algorithms differ in the way they assign patterns to different clusters. It can be simple, elongated, overlapped, or compact in nature. The three approaches are:

a) Single link: The distance between two clusters is minimum taking all pairs of patterns, with each pair having one point from both clusters. It produces loosely connected clusters which suffers from chaining effect and the clusters are not well defined i.e. few patterns that can't be specified to a single cluster are left unprocessed and are termed as noisy patterns. It is good to separate these clusters using concentric data patterns.

b) Complete link: The distance between all pairs of patterns in two clusters is taken as maximum. It produces compact clusters with well-defined boundaries [4]. However it is not as versatile as single link but is very advantageous in many pragmatic applications.

c) Minimum variance: the distance between all pairs of two clusters is taken in a way to yield minimum variance.

In all three, clusters are merged to form bigger clusters by minimum distance criteria for each pattern. However, Hierarchical clustering works on agglomerative concept and uses non-isotropic data sets.

**B. Partitioning Clustering Techniques**

Partitioning clustering obtains a single cluster structure for each pattern instead of a complete cluster of data. It is used with large data sets where adequate system memory or other resources are not available. As opposed to dendrogram representation, Partitional clustering has a advantage to be used in applications where computer representation of data is required. However, it encounters the problems of setting initial number of clusters. For optimal representation of data, combinatorial search is prohibitive due to large data size and hence, multiple run and trial method is used to obtain optimal number of clusters.

a) Squared error Algorithm: This is a most popular algorithm used for partitioning of isolated and compact form of data [5]. It calculates the squared error distance as follow:

\[
(e)^2 = \sum_{j=1}^{K} \sum_{i=1}^{n} ||x_i - c_j||^2
\]

where \( x_i \) is a \( i^{th} \) pattern belonging to \( j^{th} \) cluster and \( c_j \) is a centroids of \( j^{th} \) cluster; \( e^2 \) calculates a mean error distance from one cluster to all other nearby clusters.

---

The widely used k-means algorithm is also based on squared error algorithm. It starts with a random initial number of clusters and keeps changing the mean error and centroids based on similarity of data patterns until a convergence criteria is met. This algorithm is very suitable to use in number of applications due to its easy implementation and low complexity \[O(n)] but it is sensitive to setting initial seed. This algorithm does not reassign the patterns to new clusters with subsequent iterations otherwise squared error value ceases to change after third or fourth iteration.

Another variant of k-mean algorithm is ISODATA, it allows merging and splitting of new clusters [6] and thus works in a more dynamic manner. Some other few variants of dynamic algorithm are proposed in [7].

b) Graph-theoretic Clustering: This divisive type of clustering is based on well-known minimal spanning tree [MST] [8]. The data is arranged in a form of MST and then the edges with largest distance are cut-out and separated parts are formed into clusters. This process of forming clusters is faster than other clustering techniques. Single link hierarchical [9] approach is also based on graph theoretic clustering method.

C. Nearest Neighbor cluster

Since clustering is done according to similarity analysis of data patterns, an intuitive method is proposed based on nearest neighbor algorithm [10]. A cluster is formed and a minimum threshold distance is set initially. Then each new pattern is checked for minimum distance, if it is below threshold it is assigned to existing cluster else a new cluster is formed for farther distant points. This method insures that all nearby patterns fall in the same cluster and clusters are formed based upon density of vector points in space.

D. Fuzzy Clustering

As opposed to hard clustering where clusters are independent of each other, Fuzzy logic exploits the dependency nature between clusters [11]. Each cluster shares some similarity measure with every other cluster and is mapped by a predefined mapping function. For instance, the mapping function value varies in interval [0, 1] with nearest pattern having value 0.1 and farthest having 0.9.

The most popular fuzzy algorithm is Fuzzy c-means [FCM] algorithm [12]; it performs better than hard k-means algorithm for unsupervised and stochastic data. Fuzzy logic is good in avoiding local minima which presents in k-means algorithm but may converge to local minima's of squared error problem. Overall Fuzzy outperforms (Fig. 1) than many other popular algorithms but its performance depends on the designer's capability to select optimized membership value for clustering.

\[
\begin{align*}
X_i & \quad F_1 \\
1 & 2 \\
3 & 4 \\
5 & 6 \\
7 & 8 \\
9 & 10 \\
\end{align*}
\]

Fig 1. Fuzzy Logic membership representation

E. Representation of clusters

Whenever a data is processed or organized, the resultant effect is usually clustering. The shape of clusters affects the performance and acceptability of the proposed algorithm to a great extent. Representation of data is an important step in decision making and depends on the application of the algorithm [13].

Clusters can be defined in a form of circles with definite centroids, ellipse, dendrogram, tree or simply connected dots of points to form boundaries in case of amorphous data sets. Data abstracted from clusters should ease the complexity of algorithm and should be easy for human comprehension.

Representing data with centroids is most popular but it works well only with isotropic data. In case of elongated or non-isotropic data, collection of boundary points is most appropriate method. Also no. of data points within a cluster should also increase as complexity of the shape increases.

\[
\begin{align*}
X_i & \quad \text{F1} \\
1 & 2 \\
3 & 4 \\
5 & 6 \\
7 & 8 \\
\end{align*}
\]

Fig 2. Representing chain like data clusters

Data compression techniques are further exploited in [14]. It highlights that chain like structures as shown in Fig. 2 cannot be separated by k-means algorithm but are well suited to single link algorithm however, it is computationally expensive. The proposed solution in such a case is hybrid clustering. It exploits desirable properties from both the algorithm. Another method to simplify data reduction is to use sub-clusters within the main clusters.

F. Artificial Neural Networks for clustering

The concept of artificial neural networks is derived from nervous system and brain connection in human body. This model is being used in computers for data abstraction from last three decades and thus has developed many folds. Some basic properties neural networks possess are there vectors possess only quantitative data, have inherently parallel architecture and can learn about its weights adaptively.

Competitive algorithm (or Winner-take-all) algorithm [15] based on neural networks is very popular and exploits the concept of nervous system in human brain. In this, similar data patterns are grouped together and represented as a single neuron unit.

Another example of clustering algorithm is based on unsupervised learning using vector quantization namely linear vector quantization[LVQ] [16], Self-Organization map[SOM] [17], and adaptive resonance theory models [18]. They are single layered architectures. The underlying concept of these neural networks is simple: patterns are mapped at input and as well as at output node. The weights at both nodes are changed iteratively until a desired criterion is met.

Relationship between k-means and LVQ are discussed in [19]. Addressing some characteristics of SOM, it is most widely used algorithm for providing two dimensional clusters with multi-dimensional input data and thus, is used in vector quantization and speech recognition.

If initial weights are not chosen properly, SOM results in sub-optimal clusters and may suffers from stability issues. Some input pattern results in new output patterns with each iteration: in such cases learning rate never leads to zero with finite number of iterations. Further SOM is suitable only for hyperspherical data sets and number of output nodes are also fixed which limit the size of data clusters.

G. Evolutionary Approaches for clustering

The concept of evolutionary approaches is motivated by natural biological process of evolution i.e. survival of fittest [20]. Here, vector points are coded as chromosomes and each transformation produces a one or more output combination (called offspring) from one or more input chromosomes (called parents). The common evolutionary operators used here are selection, recombination and mutation. After obtaining different combinations, a fitness function is used to evaluate a likelihood of a set of combinations that suits best in the pre-defined environment condition and passes the criteria of survival in next generation. As perceived from biologist’s point of view, these algorithms are sufficiently complex to provide robust and powerful data clustering.

Genetic algorithm is a well-known data clustering technique used to evaluate data in highly complex environments. Typically, it uses binary strings as an input and produces output strings based on the maximum fitness value which propagates to next generation and produces more strings of its type.

Another variant of recombination used is crossover algorithm: It works in a simple manner of producing a pair of offspring from a pair of parent chromosomes by applying cross multiplication at center point. It is depicted in Fig. 3.

In mutation: single input binary string results in a output binary string by complementing one digit at some random location. For instance, 11101001 produce 11001001. In GA, both crossover and mutation are applied with a pre-defined fitness function to depict data clusters in vectored search space.

Evolutionary Programming(EP) and Evolutionary strategies (ES) can also be used in Data Clustering and are differ from GA only in the way it represents the output data. All three approaches use minimum squared error criteria as discussed in partition algorithm. Some issues related to convergence criteria are discussed in [21]. A major problem in GA [22] is its sensitivity in selecting parameters for fitness function, crossover and mutation probabilities, population size etc.

IV. COMPARATIVE ANALYSIS

A standard approach cannot be used to compare the above techniques rather a qualitative as well as quantitative approach has to be considered. At the top level of hierarchy, all the clustering approaches can be classified on the basis of search methodology. Search technique can be divided as stochastic or deterministic. Deterministic techniques guarantee an optimal partition by enumeration iterations and takes time, while stochastic gives near optimal solution in a quite less time and guarantee convergence to optimal partitions asymptotically.

Among considered techniques, only evolutionary approach is stochastic in nature and all others are deterministic. Stochastic allows permutation, has non-zero probabilities and converges to non-local optimal directions while deterministic are greedy descent in nature. Stochastic can be sequential or parallel but evolutionary are inherently parallel algorithm and can be used with parallel hardware to


186
improve systems’ speed. Evolutionary approach is globalized and population-based search technique i.e. it searches for than one solution at a time. Rests are localized search techniques and searches a single solution at a time. GA, ANN, TS [23], are all sensitive to parameter selection but evolutionary can give optimal solution with discontinuous criteria function also.

Since majority of algorithms uses squared error criteria for partitioning data clusters, they are not as flexible and reliable as hierarchical algorithms. GA and TS are better than k-means in terms of solution-quality but slower in terms of execution time [24]. GA is faster than TS in searching best solution but slower in convergence i.e. in obtaining clusters with best solution.

Only k-means and SOM can be used with large data sets and all other approaches are tested with less than 1000 data size and low dimensionality. The k-means is best because it converges to locally optimal solution as its seed could be selected properly but GA, ANN, TS, and Fuzzy are not because all these have complex learning/control parameters to be selected in advance and need large execution time for large data sets [25].

One possible solution in GA, ANN, TS and fuzzy to improve their performance is to use domain knowledge in either implicit or explicit way to guide the clustering process during execution time. But such solution produces hyper-spherical clusters and can be a limitation. Thus, hierarchical are better to use than Partitional techniques to avoid hyper-spherical clusters shape.

V. PRAGMATIC ISSUES WITH LARGE DATA SETS

With the above comparisons, following points are also necessary to elucidate in order to conclude the survey report on clustering: Conceptual knowledge, clustering descriptions, and managing large data sets. Mathematically numerical methods are used for grouping the data but using conceptual knowledge as in Fuzzy gives superior clustering results. However, patterns obtained from numerical and conceptual clustering are equivalent in representation [26].

Clusters which are described by logical expressions and probabilities can be represented in both qualitative and quantitative and are conjunctive in nature and more intuitive for human brain.

Conceptual knowledge based clustering is limited by large data sizes. Typically, 1000 no of objects can be grouped with less than 100 domain knowledge points. Divide and conquer rule is used for clustering large data sets. N number of objects can be represented as p*k where, p number of blocks are clustered with k number of patterns in each block. Then k-means algorithm can be applied in each block.

VI. CONCLUSION

Concluding the paper we observed that clustering is a subjective approach, same set of data needs to be partitioned differently for different applications. Most of the schemes for feature extraction are iterative in nature and thus can’t be used with large data sets due to computational cost. However, methods are available which do not store complete data in a computer at a time but can process one pattern at a time.

Clustering is a challenging problem and a basic step in information retrieval process. To exploit its potential, several decisions choices has to be made all depending upon the available resources and application. Arbitrary shapes, high dimensional database, and domain knowledge etc., this paper we have comparative and analytic survey on effective clustering algorithms though which the next step task could execute to satisfy aim of clustering more appropriately. The observed conclusive point is algorithm of clustering require update to form cluster

REFERENCES

analysis. Applied statistics, 54-64.


Author Details

Garima Singhal
Department of ECE, NIT Arunachal Pradesh, India
garimasinghal.888@gmail.com

Sahadev Roy
Department of ECE, NIT Arunachal Pradesh, India
sahadevroy@gmail.com