Cluster Analysis and Data Clustering

Raja’a Salih
Al Mansoor-Medical Technical Institutes
Foundation of Technical Institutes

ABSTRACT

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). The clustering problem has been addressed in many contexts and by researchers in many disciplines; this reflects its broad appeal and usefulness as one of the steps in exploratory data analysis.

However, clustering is a difficult problem combinatorially, and differences in assumptions and contexts in different communities has made the transfer of useful generic concepts and methodologies slow to occur. This paper presents an overview of pattern clustering methods from a statistical pattern recognition perspective, with a goal of providing useful advice and references to fundamental concepts accessible to the broad community of clustering practitioners. We present a taxonomy of clustering techniques, and identify cross-cutting themes and recent advances. We also describe some important applications of clustering algorithms such as image segmentation, object recognition, and information retrieval.

الخلاصة

التحليل العنقودي والتجميع العنقودي للبيانات الترتيب العنقودي هو التصنيف اللا اشراقي للانماط (الملاحظات والمشاهدات , فقرات البيانات, متجهات المعالم او المعالم الاتجاهيه,) الى مجاميع (عناقيد) . ان مساله الترتيب العنقودي للبيانات (تجميعها عنقوديا) جرى معالجتها وتداولها بعدة سياقات من قبل الباحثين في العديد من فروع المعرفة, وهذا يعكس الاستهواء الواسع والمفيد لها باعتبار انها واحدة من الخطوات في تحليل البيانات الاستكشافي.

رغم ذلك, فان الترتيب العنقودي للبيانات هو مساله صعبه من الناحية التوليفية او التجميعية والاختلافات او الفروق في الافتراضات والسياقات في العديد من المجاميع المختلفة جعلت من انتقال وتطبيق المفاهيم النوعية الشاملة المفيدة والمنهجيه مساله بطيئة الحدوث .

في هذا البحث نقدم تصورا عاما لطرائق الترتيب العنقودي للانماط من زواية تمييز الانماط احصائيا بهدف اعطاء نصيحة مفيدة ومرجعيه الى المفاهيم الاساسية المتاحة الى المجتمع الواسع من ممارسي الترتيب العنقودي في تحليل البيانات.

في هذا البحث تصنيفا لتقنيات الترتيب العنقودي وما قد استجد فيه حديثا كذلك وصف لبعض التطبيقات المهمة في خوارزميات العنقودية مثل تقطيع الصور, التعرف على الاهداف كذلك استرجاع المعلومات والبيانات.

1. INTRODUCTION

Cluster analysis is the organization of a collection of patterns (usually represented as a vector of measurements, or a point in a multidimensional space) into clusters based on similarity.

Intuitively, patterns within a valid cluster are more similar to each other than they are to a pattern belonging to a different cluster.

The variety of techniques for representing data, measuring proximity (similarity) between data elements, and grouping data elements has produced a rich and often confusing assortment of clustering methods.

It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification). In supervised classification, we are provided with a collection of *labeled* (preclassified) patterns; the problem is to label a newly encountered, yet unlabeled, pattern. Typically, the given labeled (*training*) patterns are used to learn the descriptions of classes which in turn are used to label a new pattern.

In the case of clustering, the problem is to group a given collection of unlabeled patterns into meaningful clusters. In a sense, labels are associated with clusters also, but these category labels are *data driven*; that is, they are obtained solely from the data.

The term “clustering” is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the contexts in which clustering is used.

1.1. COMPONENTS OF CLUSTERING TASK

Typical pattern clustering activity involves the following steps (Brailovsky,1991,p193)

(1) pattern representation (optionally including feature extraction and/or selection),
(2) definition of a pattern proximity measure appropriate to the data domain,
(3) clustering or grouping,
(4) data abstraction (if needed), and
(5) assessment of output (if needed).

Figure 1 depicts a typical sequencing of the first three of these steps, including a feedback path where the grouping process output could affect subsequent feature extraction and similarity computations.

Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. Some of this information may not be controllable by the practitioner .

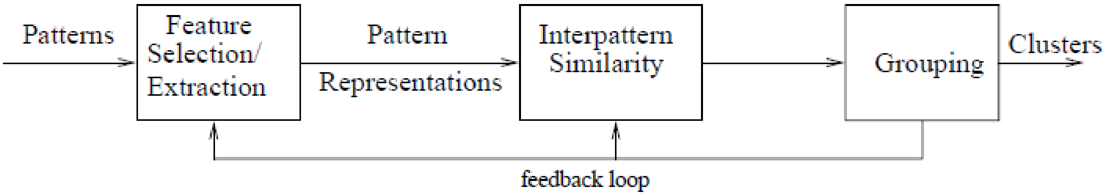

Figure 1 Stages in clustering.

Feature selection is the process of identifying the most effective subset of the original features to use in clustering. *Feature extraction* is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering. *Pattern proximity* is usually measured by a distance function defined on pairs of patterns. A variety of distance measures are in use in the various communities (Dubes,1987,p645,Cheng,1995,p895).

A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns, whereas other similarity measures can be used to characterize the conceptual similarity between patterns (Zahn,1971,p68). Distance measures are discussed in Section 4. The *grouping* step can be performed in a number of ways. The output clustering(or clusterings) can be hard (a partition of the data into groups) or fuzzy (where each pattern has a variable degree of membership in each of the output clusters). Hierarchical clustering algorithms produce a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. Partitional clustering algorithms identify the partition that optimizes (usually locally) a clustering criterion.

Additional techniques for the grouping operation include probabilistic (Jain & Flaynn,1996,p65) , and graph-theoretic (Ohler & Grey,1995,p461) clustering methods. *Data abstraction* is the process of extracting a simple and compact representation of a data set. Here, simplicity is either from the perspective of automatic analysis (so that a machine can perform further processing efficiently) or it is human-oriented (so that the representation obtained is easy to comprehend and intuitively appealing). In the clustering context, a typical data abstraction is a compact description of each cluster, usually in terms of cluster prototypes or representative patterns such as the centroid .

The study of *cluster tendency*, wherein the input data are examined to see if there is any merit to a cluster analysis prior to one being performed, is a relatively inactive research area (Rasemussen,1992,p419,Sneath,1973).

Cluster validity analysis, by contrast, is the assessment of a clustering procedure's output.

The concept of *density* clustering and a methodology for decomposition of feature spaces (Jain & Flynn,1996,p65) have also been incorporated into traditional clustering methodology, yielding a technique for extracting overlapping clusters.

Even though there is an increasing interest in the use of clustering methods in pattern recognition (Dubes,1987,p645),image processing (King,1967,p86) and information retrieval (Ward,1963,p236,Murtagh,1984,p354), clustering has a rich history in other disciplines(Brailovsky,1991,p193) such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing.

1.2. SIMILARITY MEASURES

Since similarity is fundamental to the definition of a cluster, a measure of the similarity between two patterns drawn from the same feature space is essential to most clustering procedures. Because of the variety of feature types and scales, the distance measure (or measures) must be chosen carefully. It is most common to calculate the *dissimilarity* between two patterns using a distance measure defined on the feature space.

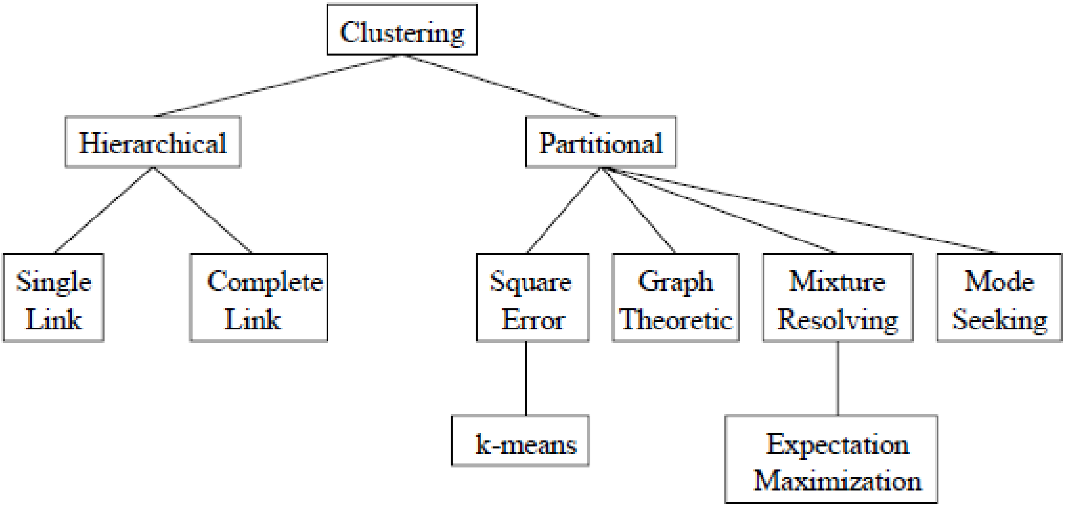

Figure 2 A taxonomy of clustering approaches.

The most popular metric for continuous features is the *Euclidean distance*

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d}(x_{i,k} - x_{j,k})^2\right)^{1/2}$$

$$= \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

which is a special case of the Minkowski metric

$$d_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d}|x_{i,k} - x_{j,k}|^p\right)^{1/p}$$

$$= \|\mathbf{x}_i - \mathbf{x}_j\|_p.$$

The Euclidean distance has an intuitive appeal as it is commonly used to evaluate the proximity of objects in two or three-dimensional space. It works well when a data set has “compact” or “isolated” clusters .The drawback to direct use of the Minkowski metrics is the tendency of the largest-scaled feature to dominate the others. Solutions to this problem include normalization of the continuous features (to a common range or variance) or other weighting schemes. Linear correlation among features can also distort distance measures; this distortion can be alleviated by applying a whitening transformation to the data or by using the squared Mahalanobis distance

$$d_M(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)\Sigma^{-1}(\mathbf{x}_i - \mathbf{x}_j)^T,$$

where the patterns **x***i* and **x***j* are assumed to be row vectors

2.CLUSTERING TECHNIQUES

Different approaches to clustering data can be described with the help of the hierarchy shown in Figure 2 (other taxonometric representations of clustering methodology are possible; ours is based on the discussion in Jain and Dubes). At the top level, there is a distinction between hierarchical and partitional approaches (hierarchical methods produce a nested series of partitions, while partitional methods produce only one).
Most hierarchical clustering algorithms are variants of the single-link (Baeza-yates,1992,p13), complete-link (Nagy,1968,p836), and minimum-variance(Fisher,1993,p51,Salton,1991,p974) algorithms. Of these, the single-link and complete-link algorithms are most popular. These two algorithms differ in the way they characterize the similarity between a pair of clusters. In the single-link method, the distance between two clusters is the *minimum* of the distances between all pairs of patterns drawn from the two clusters (one pattern from the first cluster, the other from the second).

In the complete-link algorithm, the distance between two clusters is the *maximum* of all pair wise distances between patterns in the two clusters. In either case, two clusters are merged to form a larger cluster based on minimum distance criteria. The complete-link algorithm produces tightly bound or compact clusters (Jain & dubes,1988). The single-link algorithm, by contrast, suffers from a chaining effect (Anderberg,1973,diday & Simon,1976,p47). It has a tendency to produce clusters that are straggly or elongated .

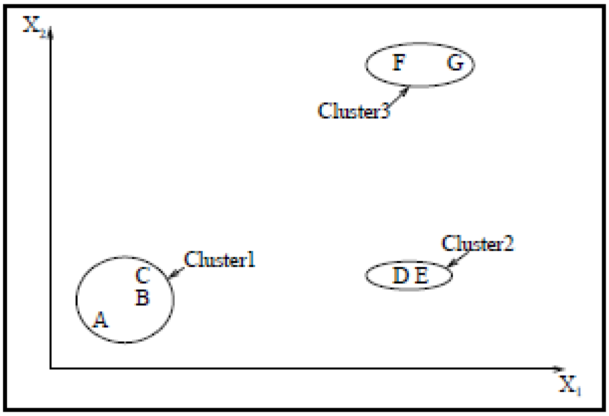

Figure 3 points falling in three clusters

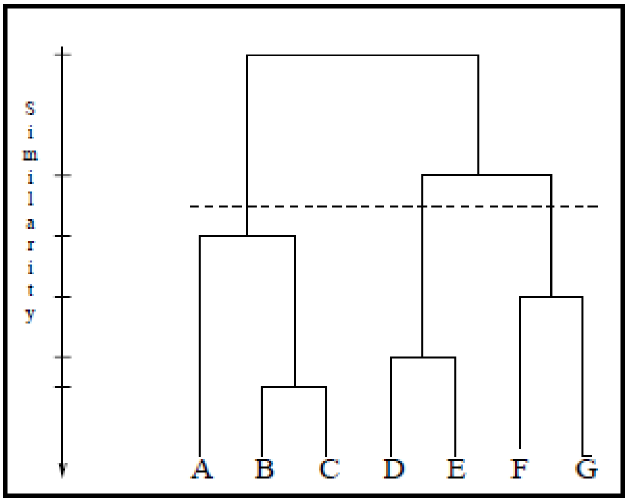

Figure 4 The dendrogram obtained using the single – link algorithm

3. CLUSTERING GROUPING

Clustering is a process of grouping data items based on a measure of similarity. It is also a subjective process ; the same set of data items often needs to be partitioned differently for different applications. This subjectivity makes the process of clustering difficult. This is because a single algorithm or approach is not adequate to solve every clustering problem. A possible solution lies in reflecting this subjectivity in the form of knowledge. This knowledge is used either implicitly or explicitly in one or more phases of clustering.

Knowledge-based clustering algorithms use domain knowledge explicitly. The most challenging step in clustering is feature extraction or pattern representation. Pattern recognition researchers conveniently avoid this step by assuming that the pattern representations are available as input to the clustering algorithm.

In small size data sets, pattern representations can be obtained based on previous experience of the user with the problem. However, in the case of large data sets, it is difficult for the user to keep track of the importance of each feature in clustering. A solution is to make as many measurements on the patterns as possible and use them in pattern representation. But it is not possible to use a large collection of measurements directly in clustering because of computational costs. So several feature extraction/selection approaches have been designed to obtain linear or nonlinear combinations of these measurements which can be used to represent patterns. Most of the schemes proposed for feature extraction/selection are typically iterative in nature and cannot be used on large data sets due to prohibitive computational costs. The second step in clustering is similarity computation.

A variety of schemes have been used to compute similarity between two patterns. They use knowledge either implicitly or explicitly. Most of the knowledge-based clustering algorithms use explicit knowledge in similarity computation.

However, if patterns are not represented using proper features, then it is not possible to get a meaningful partition irrespective of the quality and quantity of knowledge used in similarity computation.

There is no universally acceptable scheme for computing similarity between patterns represented using a mixture of both qualitative and quantitative features. Dissimilarity between a pair of patterns is represented using a distance measure that may or may not be a metric. The next step in clustering is the grouping step. There are broadly two grouping schemes: hierarchical and partitional schemes. The hierarchical schemes are more versatile, and the partitional schemes are less expensive. The partitional algorithms aim at maximizing the squared error criterion function. Motivated by the failure of the squared error partitional clustering algorithms in finding the optimal solution to this problem, a large collection of approaches have been proposed and used to obtain the global optimal solution to this problem. However, these schemes are computationally prohibitive on large data sets. ANN-based clustering schemes are neural implementations of the clustering algorithms, and they share the undesired properties of these algorithms. However, ANNs have the capability to automatically normalize the data and extract features. An important observation is that even if a scheme can find the optimal solution to the squared error partitioning problem, it may still fall short of the requirements because of the possible non-isotropic nature of the clusters.

In some applications, for example in document retrieval, it may be useful to have a clustering that is not a partition. This means clusters are overlapping. Fuzzy clustering and functional clustering are ideally suited for this purpose. Also, fuzzy clustering algorithms can handle mixed data types. However, a major problem with fuzzy clustering is that it is difficult to obtain the membership values.

A general approach may not work because of the subjective nature of clustering. It is required to represent clusters obtained in a suitable form to help the decision maker. Knowledge-based clustering schemes generate intuitively appealing descriptions of clusters. They can be used even when the patterns are represented using a combination of qualitative and quantitative features, provided that knowledge linking a concept and the mixed features are available. However, implementations of the conceptual clustering schemes are computationally expensive and are not suitable for grouping large data sets.

The *k*-means algorithm and its neural implementation, the Kohonen net, are most successfully used on large data sets. This is because *k*-means algorithm is simple to implement and computationally attractive because of its linear time complexity. However, it is not feasible to use even this linear time algorithm on large data sets. Incremental algorithms like leader and its neural implementation, the ART network, can be used to cluster large data sets. But they tend to be order-dependent. *Divide and conquer* is a heuristic that has been rightly exploited by computer algorithm designers to reduce computational costs. However, it should be judiciously used in clustering to achieve meaningful results.

In summary, clustering is an interesting, useful, and challenging problem. It has great potential in applications like object recognition, image segmentation, and information filtering and retrieval. However, it is possible to exploit this potential only after making several design choices carefully.

4. .Information Theory bases of clustering algorithm :

Clustering is a problem of great practical importance in numerous applications. The problem of clustering becomes more challenging when the data is categorical, that is, when there is no inherent distance measure between data values. This is often the case in many domains. Without a clear measure of distance between data values, it is unclear how to define a quality measure for categorical clustering. To do this, we employ mutual information, a measure from information theory. A good
clustering is one where the clusters are informative about the data objects they contain. Since data objects are expressed in terms of attribute values, we require that the clusters convey information about the attribute values of the objects in the clusters. The quality measure of the clustering is then the mutual information of the clusters and the attribute values.

On the basis of the information theory , Let T denote a discrete random variable that takes values over the set T, and let p(t) denote the probability mass function of T. The entropy H(T) of variable T is defined by :

$$H(T) = -\sum_{t\text{€}T} p(t)\log p(t)$$

let T and A be two random variables that range over sets T and A respectively, and let p(t,a) denote their joint distribution and p(a\t) be the conditional distribution of A given T. Then conditional entropy H(A\T) is
defined as :

$$\begin{aligned} H(A\backslash T) &= \sum_{t\text{€}T} p(t)H(A\backslash T=t) \\ &= -\sum_{t\text{€}T} p(t) \sum_{a\text{€}A} p(a\backslash t)\log p(a\backslash t) \end{aligned}$$

Given T and A, the mutual information, I(T;A), quantifies the amount of information that the variables hold about each other. The mutual information between two variables is the amount of uncertainty (entropy) in one variable that is removed by knowledge of the value of the other one ,then :

$$\begin{aligned} I(T;A) &= \sum_{t\text{€}T} \sum_{a\text{€}A} p(a,t)\log \frac{p(t,a)}{p(t)p(a)} \\ &= \sum_{t\text{€}T} p(t) \sum_{a\text{€}A} p(a\backslash t)\log \frac{p(t\backslash a)}{p(a)} \\ &= H(t) - H(t\backslash a) = H(a) - H(a\backslash t) \end{aligned}$$

Thus , Mutual information is symmetric, non-negative and equals zero if and only if T and A are independent.

5. APPLICATIONS

There are several applications where decision making and exploratory pattern analysis have to be performed on large data sets. For example, in document retrieval, a set of relevant documents has to be found among several millions of documents of dimensionality of more than 1000. It is possible to handle these problems if some useful abstraction of the data is obtained and is used in decision making, rather than directly using the entire data set.

By *data abstraction*, we mean a simple and compact representation of the data. This simplicity helps the machine in efficient processing or a human in comprehending the structure in data easily. Clustering algorithms are ideally suited for achieving data abstraction. In this paper, we have examined various steps in clustering:

(1) pattern representation,
(2) similarity computation,
(3) grouping process, and
(4) cluster representation.

Clustering is applied to statistical , neural, evolutionary, and knowledge-based approaches. The most important four applications of clustering are:

(1) image segmentation,
(2) object recognition,
(3)document retrieval, and
(4) data mining.

6. CONCLUSION

Clustering algorithms have been used in a large variety of applications (Murtagh,1984,p345,King,1967,p86,Sneath,1973,michalski,1983,396). we describe several applications where clustering has been employed as an essential step.

These areas are:

(1) image segmentation
(2) object and character recognition,
(3) document retrieval, and
(4) data mining

REFERENCES:

1. ANDERBERG, M. R.. *Cluster Analysis for Applications*. Academic Press, Inc., New York, NY. 1973
2. BAEZA-YATES, R. A.. Introduction to data structures and algorithms related to information retrieval. In *Information Retrieval:Data Structures and Algorithms*, W. [1] B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc.,Upper Saddle River, NJ, [1] 13–27.1992
3. BRAILOVSKY, V. L.. A probabilistic approach to clustering. *Pattern Recogn.Lett.12*, 4, 193–198.Apr.1991
4. CHENG, C. H.. A branch-and-bound clustering algorithm. *IEEE Trans. Syst.Man Cybern. 25*, 895–898.1995
5. DIDAY, E. AND SIMON, J. C.. Clustering analysis. In *Digital Pattern Recognition*, K.S. Fu, Ed. Springer-Verlag, Secaucus, NJ,47–94.1976
6. DUBES, R. C.. How many clusters are best?—an experiment. *Pattern Recogn.20*, 6(Nov. 1, 1987), 645–663.1987
7. FISHER, D., XU, L., CARNES, R., RICH, Y., FENVES, S.J., CHEN, J., SHIAVI, R., BISWAS, G., AND WEINBERG,J.. Applying AI clustering to engineering tasks. *IEEE Expert 8*, 51–60.1993
8. JAIN, A. K. AND DUBES, R. C.. *Algorithms for Clustering Data*. Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River,NJ.1988.
9. JAIN, A. K. AND FLYNN, P. J.. Image segmentation using clustering. In*Advances in Image Understanding: A Festschrift for Azriel Rosenfeld*, N. Ahuja and K. Bowyer, Eds,IEEE Press, Piscataway, NJ, 65–83.1996
10. KING, B.. Step-wise clustering procedures.*J. Am. Stat. Assoc. 69*, 86–101.1967
11. MICHALSKI, R., STEPP, R. E., AND DIDAY,E.. Automated construction of classifications:conceptual clustering versus numerical taxonomy. *IEEE Trans.Pattern Anal.Mach. Intell. PAMI-5*, 5 (Sept.), 396–409.1983
12. MURTAGH, F.. A survey of recent advances in hierarchical clustering algorithms which use cluster centers. *Comput. J. 26*, 354–359.1984
13. NAGY, G.. State of the art in pattern recognition. *Proc. IEEE 56*, 836–862.1968
14. OEHLER, K. L. AND GRAY, R. M..Combining image compression and classification using vector quantization. *IEEE Trans.Pattern Anal. Mach.Intell. 17*, 461–473.1995
15. RASMUSSEN, E.. Clustering algorithms. In *Information Retrieval: DataStructures and Algorithms*, W. B. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Inc., Upper Saddle River, NJ, 419–442.1992
16. SALTON, G.. Developments in automatic text retrieval. *Science 253*, 974–980.1991
17. SNEATH, P. H. A. AND SOKAL, R. R.. *Numerical Taxonomy*. Freeman, London,UK.1973
18. WARD, J. H. JR.. Hierarchical grouping to optimize an objective function. *J.Am. Stat.Assoc. 58*, 236–244.1963
19. ZAHN, C. T.. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput. C-20* (Apr.), 68–86.1971