IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Clustering Methods in Data Mining Techniques- A Review

M. Latha [1], Dr. K. Subramanian, M.Sc., M.Phil., Ph.D., [2]

Research Scholar, JJ College of Arts and Science (Autonomous), Pudukkottai [1]

Assistant Professor, H.H. The Rajah's College (Autonomous), Pudukkottai [2]

Abstract: Clustering is an unsupervised classification of the cluster pattern observations in the group of data elements or eigenvectors. The caking problem is solved in many cases and in many disciplines by researchers. This reflects its broad appeal and practicality as a step in exploratory data analysis. However, grouping is a difficult combination of assumptions in different communities and contextual differences have made useful general concepts and methods of slow turn. This article introduces the basic concepts of statistical pattern recognition and community access to a wider community of professionals, providing an overview of the methods of grouping patterns of useful tips and reference targets. We propose clustering techniques for categorizing and identifying cross-problems and recent developments. Some important applications of clustering algorithms such as image segmentation target recognition and information retrieval are also described.

Keywords: Clustering Algorithm, Data Analysis, Feature Selection, Feature Extraction.

I. INTRODUCTION

A. Motivation

Data analysis underlies many applications, either at the design stage or as part of their on-line operation. Data analysis programs can be used for dichotomous or exploration conrmatorios, based on the availability of appropriate models for data sources, but in both types of procedures (whether training or making assumptions) are based on a hypothesis model or Ii) by analyzing (i) the grouping or classification of goodwill measures. Clustering analysis is a collection of patterns (usually expressed as a vector of measurements, or a point in a multidimensional space) that is inserted according to the organization of similar groups. Intuitively, the patterns within the active set are more similar to each other than they belong to a different group of patterns. Cluster analysis is as old as human life and has its roots in many fields, such as statistics, machine learning, biology, and artificial intelligence. Therefore, cluster analysis is called in different areas such as analysis Q, typology, clustering, numerical classification, data segmentation, supervised learning, data visualization and observation learning [1].

It is important to understand the difference between clustering (unsupervised classification) and discriminant analysis (supervised classification). In supervised classification, we provide a set of marker patterns (vaccinations) and the problem is the pattern of the newly discovered label, but not the marker. In the case of the group, the problem is that the clusters in the given set of patterns [2] are not labeled with clusters. Clustering is useful in several exploratory analyzes of patterns, clusters, decision making and machine learning situations, including data mining, document retrieval, image segmentation and pattern classification.

However, in many of these problems, few a priori information (eg, statistical models) are provided on the data and the decision maker should make assumptions about the data as few as possible [3]. It is in these limitations that clustering analysis is particularly suited to the exploration of the relationships between data points, making their structural (perhaps preliminary) assessments.

The term & quot; cluster & quot; is used in several research communities to describe methods, and data packets are not labeled. These communities have components for clustering and use different terms and assumptions in the context of the group. Thus, we are confronted with a range of dilemmas about the survey [4]. The production of truly comprehensive surveys will be entrusted with a great deal of arduous task in this regard. The accessibility of the survey may also be problematic because of the need to reconcile very different vocabularies and assumptions about grouping in different communities. The purpose of this paper is to summarize the basic concepts and techniques of clustering large clusters of statistical and decision theory roots. Where appropriate, refer to the key concepts and techniques learned from cluster methodological machines and from other communities.

B. Components of a Clustering Task

Typical model cluster activities include the following steps [5].

(1) mode representation (optionally including selection of extraction and / or properties)

(2) a definition of an appropriate measure approaching to the domain pattern data, groups

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

(3) data abstraction (if necessary),
(4) And assessment of output (if needed).
The typical sequence of the first three of these steps, including the output of the clustering process, may affect the subsequent feature extraction and similarity computation of the feedback path [6].

The pattern representation refers to the number of classes, the number of available modes, and the number, types, and sizes of features of the available clustering algorithm. Some information may not be controlled by a professional. Feature selection is the process of determining the most efficient atomic set (Figure 1)

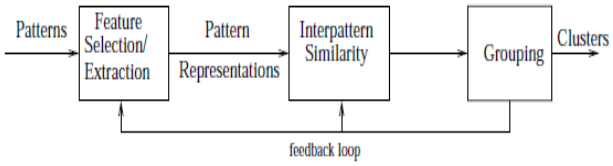

Fig.1 Components of clustering task

Feature extraction is one or more transformations of the input characteristics used to produce new good features. One or both of these techniques may be used to obtain an appropriate set of characteristics for use in the packet.

The vicinity of the pattern is usually signaled by the distance function defined for the pattern. The various measures of distance are used in the various communities [7] [8]. A simple measurement distance as a Euclidean distance can often be used to reflect the difference between the two patterns, and other similarity measures can be used to characterize the conceptual similarity between patterns.

Data abstraction is the process of extracting a simple and compact representation of a data set. In this case, or simply from the perspective of automatic analysis (for a machine that allows efficient and efficient processing) is directed to the person (the resulting representation is easy to understand and intuitively appealing).

C. History of Clustering
Clustering has a rich history in other disciplines [5], such as biology, psychiatry, psychology, archeology, geology, geography and marketing. Other more or less synonyms of grouping include unsupervised learning, numerical classification, vector quantization, and observational learning [9].
The point of the graph [10] The spatial analysis of the domain also involves cluster analysis. The importance of grouping and the interdisciplinary nature of the literature through its vastness are evident.

They have already published several books on clustering, plus some useful inienciales for reviewing projects. The various clustering algorithms in the 1978 cluster state survey were compared to establish the minimum spanning tree and short extension paths were done.

Report grouping. Several program combinations were optimized on the basis of experimental comparisons that have been reported in inches.

II. DEFINITIONS AND NOTATION

The following terms and notation will be used throughout this paper.
A pattern (or entity vector, observation or datum) x is a unique data element used by the clustering algorithm. Normally it consists of a vector of measurements d: x = (x1 d).
A single scalar component of the pattern X is called a function (or tribute), and D is the dimension of the spatial pattern or pattern.

The clustering technique attempts to group the patterns so that the gradations thus obtained are reflected in the different pattern patterns represented by the group pattern.
The hard clustering technique class assigns classes to each patter to determine its class. All labels for the set of patterns X of the set are L = {L 1 ... LN}, and L i {L1, ..., K}, where k is the number of clusters.

The program assigns a fuzzy membership to each input mode ti fractional degree of the degree of each packet j component fixed.
Distance measurement (close-to-specialization) is a metric (or quasi-metric) that is used to quantify similar feature space of a pattern.

III.PATTERN REPRESENTATION, FEATURE SELECTION, AND FEATURE EXTRACTION

Pattern generation process is often not directly controlled. The role of the user of the process represented by the schema is to collect data and hypothetical data, optionally to complete the selection and the extraction of features and the subsequent design of the elements for grouping. Because of the representative difficulty surrounding the pattern, it may be convenient to assume that the representation of the pattern is provided before clustering [11]. However, the existing functions of meticulous detection and any (even the simplest) convertible can produce significant improvements in packetization. A well-represented pattern tends to produce an easy-to-understand grouping. The poor representation of a pattern can produce complex packets whose real structure is difficult or impossible to discern. . The 2D features divided in the space are arranged in groups of curves at a substantially constant distance from the source. If the Cartesian coordinates are chosen to represent the pattern, it is likely that many of the clustering algorithm fragments are clustered into two or more groups, and it is not compact. If, however, for a polar coordinate representation for a compact, the radius coordinates show a close-up grouping and it is possible to easily obtain a solution in a group Patterns can measure physical objects (such as chairs) or an abstract concept (for example, write styles).

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

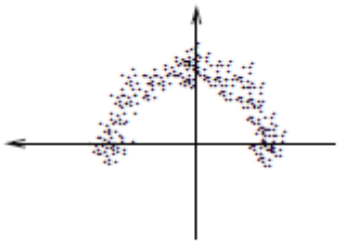

Fig.2 Pattern Representation of Clustering

As mentioned above, the pattern is usually represented as a multidimensional vector (Figure 2), where each dimension is a unique feature [12]. These features may be quantitative or qualitative. For example, if the weight and color are used for both functions, then (20 black) it is represented by a 20-unit weight black object.

The purpose of a symbol is defined by a logical combination of events. These event links, where the features can take one or more values, and are not necessarily limited to values and properties in all objects of the same set of features.

It is often valuable to isolate only the most descriptive and discriminatory features of the input set and to use these features only in subsequent analyzes. The feature selection technique determines a subset of existing functions for later use, and the feature extraction technique calculates the original series of new functions. In any case, our goal is to improve the effectiveness of classification and / or computational efficiency. Feature selection is a problem of statistical pattern recognition and exploration, however, the process of feature selection must be temporary in the context of the group (ie, without class labels for the pattern), involving several sub-sets of trials and there, The resulting [6] selects the wrong process.

IV. SIMILARITY MEASURE

A set of self-similarities is indispensable for the definition of a measure of similarity between two patterns in the same extracted feature space which is indispensable for most program groups. Due to various functions and scales, distance measurements (or measures) should be chosen with caution [13]. We will focus on patterns, which are characterized by all the successive well known distance metrics.

$$d_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2 \right)^{1/2} = \|\mathbf{x}_i - \mathbf{x}_j\|_2,$$

which is a special case (p=2) of the Minkowski metric

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

Euclidean distance is intuitively appealing because it is usually used to assess the proximity of space objects in two or three dimensions. It works well when a data set is compact or isolated from a cluster. The drawback of the direct use of the Minkowski metric is the tendency to dominate the larger scale of features [14].

V. CLUSTERING TECHNIQUES

Different approaches to clustering data can be described with the help of the hierarchical structure shown in the diagram (other taxonométricas negotiation grouping approaches are possible, we are based on the discussion [15]). At the top level, there are hierarchical and hierarchical methods of distinction (hierarchical methods produce a series of nested partitions, whereas a partitioning method produces only one). The classification shown in the figure should be supplemented by a discussion that is able to (in principle) affect all the different methods, regardless of their crossover problem in the classification position. All the basic clustering algorithms can be divided into two categories (Figure 3) and hierarchical partitioning on the basis of the generated cluster attributes.

Agglomerative vs. divisive: This aspect is related to the structure and algorithmic operation. A cohesive approach starts in different clusters (single) per mode, and then organizes the clumps until the stop criterion is met. A split method is used, all the patterns start in a group until the stop criterion is met.

Monothetic vs. polythetic: This refers to a function that is used continuously or simultaneously during clustering. Most of the algorithms are politíticos, ie, all the functions in the calculation of the distance between the patterns, and decide based on these distances. A simple algorithm in turn considers the properties of the monotético reports to partition the given set of patterns.

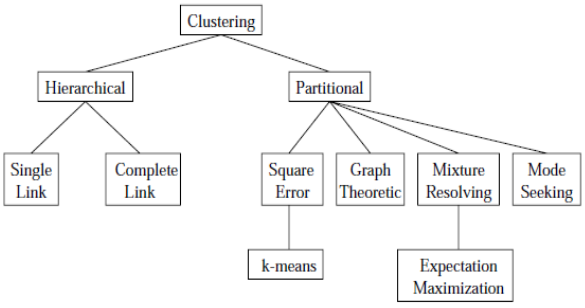

Fig.3 A taxonomy of clustering approaches.

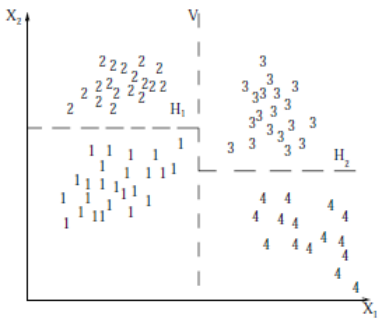

Fig.4 Monothetic partitional clustering

DOI 10.17148/IJARCCE.2016.51186

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Hard vs. fuzzy: The hard-clustering algorithm runs and outputs a group for each mode during the output. The fuzzy clustering method distributes the membership degrees in each of the input patterns in the plurality of clusters. Fuzzy clustering can be converted into a hard cluster by assigning each pattern to the cluster with the highest metric belonging to it.

Deterministic vs. stochastic: This problem is most relevant and is intended to optimize the way quadratic function errors are partitioned. This optimization can be done using conventional techniques or through a random search by the state space of all possible markers.

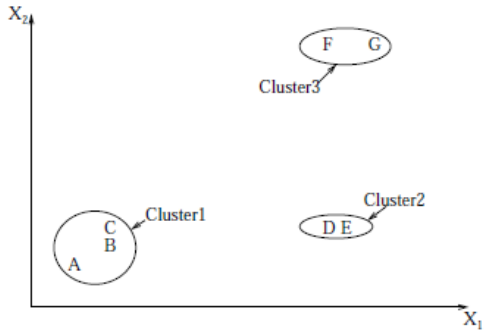

Fig.5 Points falling in three clusters

A. Hierarchical Clustering Algorithms
The running hierarchical clustering algorithm is shown using a two-dimensional dataset. The hierarchical algorithm generates a tree-like representation of the clustered nesting pattern and the level of similarity, where the cluster changes. Dendrograms can be divided into different levels to produce different data groups.

Most clustering algorithms are algorithmic single-link, full-link and minimal variance variants. Among them, only the link algorithm and link are all the most popular. These algorithms differ in the manner in which the similarity between pairs of clusters is characterized. In the single-link method, the distance between two groups is the smallest of all pairs extracted from the two sets of distances between the patterns (the first pattern group and the other second). In the full linkage algorithm, the distance between the two groups is the largest of the two groups [14] in the pattern between all pairs of distances. In both cases, the two groups are grouped to form a larger group based on the minimum distance condition. The algorithm generates full links resolutely.

a. Agglomerative Single-Link Clustering Algorithm
Place each pattern in its own group. Creates an interpattern of distances in the list of all disordered pairs of different patterns and sorting the list in ascending order.
By way of an ordered list of step steps, each value differs from the graph type of the dk, where the double-mode approximation DK is formed by the edge connections of the graph. If all employers are members of a connected graph, stop. Otherwise, repeat this step.
The output of the algorithm is a hierarchical graphic hierarchy that can be cut into distinct levels in the corresponding graphic composition by simply identifying the desired level of a partition (group) of connected components

b. Agglomerative Complete-Link Clustering Algorithm
Put each pattern in its own cluster. Constructs an inter-mode distance list for all disparate pattern pairs and sorts the list in ascending order.
By way of an ordered list of step steps, each value differs from the graph type of the dk, where the double-mode approximation DK is formed by the edge connections of the graph. If all employers are fully connected to the members of the diagram, stop.
The output of the algorithm is a hierarchical graph hierarchy that can be cut to dissimilarity to form the desired level of the partition (packet) identified by the component of the corresponding fully connected graph.

c. Hierarchical Agglomerative Clustering Algorithm
(1) Calculating a proximity matrix comprising the distance between each pair of patterns. Treat each pattern as a group.
(2) Find the closest pair of clusters using proximity matrices. One group combined these two groups. The similarity matrix is updated to reflect this merging operation.

B. Partitional Algorithms
Divide the data obtained by the clustering algorithm, not a partition of a grouping structure, such as clustering by hierarchical techniques. Part of the approach involved in large data sets where one of the tree-building constructs is computationally prohibitive is the application of advantages.

The techniques in the partitioning typically produce cluster-optimized locally defined standard functions (in a subset of schemas), or global (defined in all modes). The combinatorial search for all possible LA-belings for the standard optimal value is clearly prohibitive. In practice, therefore, the algorithm typically runs several different start states and gets the best configuration for all performed packets for output.

a. Squared Error Algorithms.
The more intuitive methods and common clustering techniques particional are squared error criteria, which tend to work well with separated and compact clusters.
The K-means algorithm is the simplest and most commonly used algorithm, using the mean-squared error criterion. It continues with a similar initial redistribution pattern cluster between the pattern and the cluster center, with a random initial partition and until the convergence criterion is satisfied.
The K-means algorithm is popular because it is easy to implement and its time complexity is O (n), where n is the number of patterns.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

A major problem with this algorithm is that it is sensitive to the selection of the initial segmentation and if the initial partition is not properly chosen to converge to the local minimum criterion function value.

K-Means Clustering Algorithm
(1) Select the k-cluster center matching random selection k-type or the hyper-volume k-random point containing the set of patterns.
(2) Each pattern to the nearest cluster.
(3) Recalculation using the current cluster association clustering center.
If the convergence criteria are not satisfied, go to step 2. The typical convergence criteria are: the pattern, with no (or minimal) redistribution of the new center group, or a minimum of the mean square error.

C. Mixture Resolving and Mode-Seeking Algorithms
The methods to solve the hybrid clustering analysis have been solved in many ways. The basic assumption is that the pattern is drawn from a set of several distributions, and our goal is to identify the parameters of each, and (perhaps) the number. The most work in this respect has assumed that the individual components of the density of the mixture are Gaussian, and in this case the individual Gaussian parameters must be estimated by this method.
It has been applied to the problem of parameter estimation algorithm expectation maximization (EM) (general problem of missing data in maximum likelihood algorithm).

In the parametric density of the electromagnetic frames of the components they are unknown because of the mixed parameters, and these slave models are calculated. The EM process, initially estimated by the vector PA starts rameter and iterates back against the parameter vector to generate the density of the mixed mode. The re-scoring pattern is then used to update the parameter estimates.

D. Nearest Neighbor Clustering
Since proximity plays a key role in the intuitive concept of our cluster, the distance between nearest neighbors can serve as a basis for grouping procedures. An iterative process, unmarked to propose each group assignment in its nearest neighbor graphics pattern, but the distance of the labeled neighbor is below the threshold. This process continues until all the patterns have no tags or other markup occurs. The values of the neighborhood (described above in the context of the calculated distance) may also be used for clusters of neighbors that grow.

E. Fuzzy Clustering
Fuzzy clustering extends this concept to use a member function for each pattern associated with each group. The output of these algorithms is a grouping, but not a partition. The most popular algorithm, fuzzy clustering algorithm is Fuzzy C-Means (FCM). Although it is better than k-means to avoid local minima, FCM can converge to the local minimum variance criterion. The design of the membership function is the most important problem in fuzzy clustering. Various options include decomposition based on similarity and clustering centers. The generalization of the FCM algorithm presents a family of objective functions. The algorithm for fuzzy C-shell and adaptive mutation detection for circular and elliptical borders (Fig. 6) is presented.

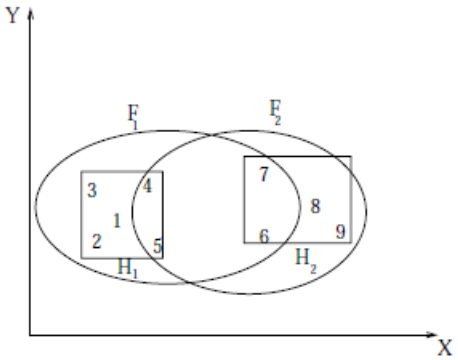

Fig.6 Fuzzy clusters.

F. Artificial Neural Networks for Clustering
Artificial neural networks (ANNs) are motivated by biological neural networks [11]. Artificial neural networks have been widely used over the past three decades for classification and clustering. Some of the features of RNA that are important in the grouping of patterns are:

(1) ANNs process numerical vectors and so require patterns to be represented using quantitative features only.
(2) ANNs are inherently parallel and distributed processing architectures.
(3) ANNs may learn their interconnection weights adaptively. More specifically, they can act as pattern normalized and feature selectors by appropriate selection of weights.

VI. APPLICATIONS

Clustering algorithms have been used in a variety of applications. In this section, we describe several applications where clustering has been used as a basic step. These areas are: (1) image segmentation, (2) recognition of objects and characters, (3) file retrieval and (4) data mining.

A. Image Segmentation Using Clustering
Image segmentation is a key component in many computer vision applications and can be viewed as a problem of clustering. The system (s) that render the image analysis (s) partition (s) the image (s) to a large extent depending on the scene being detected, the geometry, construction, and sensor of the image used to reverse-form the digital image and Scenes in the system.

The applicability of the method of segmentation image segmentation has been identified for more than thirty years and the underlying paradigm of initial efforts to exploit is

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

still present. The recurring theme in each image intensity and position of its own pixel consists of two functional positions (pixels) of the defined feature vector.

a. Segmentation
Image segmentation is typically referred to as a detailed partitioning of an input image into a plurality of regions, each of which is considered to be HO-geneous with respect to some image characteristics (e.g., intensity, color or texture) of interest.
Many Segmentors have been used for spectral measurements (eg, multispectral scanners used in remote sensing) and spatial (position-based pixels in the image plane). Therefore, the concept of directly corresponding to our pattern is measured at each pixel.

b. Image Segmentation Via Clustering.
Local entities implement grouped records to segment image gray levels. This paper emphasizes the proper selection of the characteristics of the method of each pixel, rather than the grouping, and proposes a group-based use of the image plane (spatial information) for dividing the coordinates used as additional features.
An algorithm for caking adhesive fuses (at each step) produces two sets of total minimum Mahalanobis distances between groups. The same framework is applied to texture image segmentation, but these image polynomial patterns are not sufficient for a parameterized model MRF field assumption.

B. Object and Character Recognition
a. Object Recognition**.**
Use the cluster for the purpose of grouping the data objects identified by the 3D object view objects described in the group. The term refers to an image that does not complete the field of view of an object obtained from any arbitrary point. The method used considers the problem of system object recognition for the associated view (or centered view). Each object that is recognized is represented in an object-wide image library. In this work, the identified classes are grouped and the remainder of this section will introduce the technique.
The classes of objects & apos; landscapes are grouped into similarities based on the spectral characteristics of the way. Each input image of the object is viewed in isolation to produce a representation of the view feature vector. The eigenvector contains a normalized spectral distribution of the object graph, the first ten centroids of height (H). The spectral shape of the object view is obtained from the histogram index values of its range data (the value of the curvature associated with the surface), and accumulates all the pixels of the objects falling in each compartment.

b. Clustering Views
An image database containing a range object 3200 containing 10 objects of different views 320 is used. The view of the range of images 320 is composed of possible points of the synthesized object (as determined by the use of the subdivision of the icosahedral spherical viewpoint).

c. Character Recognition.
Clusters used to recognize handwritten text semantics for freelance handwriting recognition. The success of a handwriting recognition system is critical for potential users to rely on its acceptance. The writer-dependent system provides a higher level of recognition accuracy for the independent writer system, but requires a large amount of training data. A separate system of writers, on the other hand, must be able to recognize a variety of writing styles to suit individual users. If increased writing style changes to be captured by the system. Using the Using the distances calculated in this manner, close arrays are constructed for each digital class (i.e., 0 to 9). Each matrix measures the in-class distance of a particular class. The numbers of the specific class are grouped in an attempt to create a small number of prototypes.

The distances calculated in this manner, the array of numbers for each category (i.e., 0 to 9), are approximated. Each matrix measures the in-class distance of a particular class. The numbers of the specific class are grouped in an attempt to create a small number of prototypes.
CLUSTER attempts to produce an optimal grouping for each K, where K is the number of clusters in which they are to divide the data. As expected, the mean square error (MSE) monotonically decreases to K with one of the \ best values of K through the contrast in the MSE K. The pattern recognition "knee" selection
When a set of numbers is represented by a single prototype, the number of the closest to the center of the set is used to obtain the best online identification result. Using this scheme, the correct recognition rate of 99.33% is obtained

d. Information Retrieval
Information retrieval (IR) is concerned with automatic storage and retrieval of documents.
Libraries use the United States Congress Classification (LCC) program library efficient storage and book retrieval. The architecture consists of classes LCC to Z [118], used to feature books from different disciplines. For example, the label Q corresponds to the Book Science field and the QA sub to the Math. QA76 to QA76.8 tags are used to categorize related computer and computer books and other fields.

There are a number of issues related to the use of the LCC Plan Classification. Some of them are listed below:
(1)When a user is looking for a topic in a book he is interested in dealing with, the number of LCCs alone may not be available for all related books. This is because enough information is usually entered into the database to assign a category number to a book or subject category that does not contain all the topics covered by a book. In order to illustrate this, consider the number of the book `algorithm group LCC'QA 278. J35 '. In this issue of LCC, QA 278 corresponds to the topic "Cluster Analysis", which is the name of the first author and 35 is the serial number assigned by the Library of Congress.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

This book is provided by the publisher (this category is usually signed up for clustering analysis, data processing and algoritmos. Hay in this book, dealing with computer vision processing involves a chapter on database and image segmentation imágenes. Así, the user searches computer vision literature, In particular, image segmentation cannot use the digital LCC or subject category search database to access the book that is provided in the database.

There is an inherent problem with the distribution of digital books in a rapidly developing field of LCC. For example, consider the area of the neural network. Initially, the LCC program class "QP" to mark books and conference proceedings in the region.

Specifying a number to a new book is a tricky question. A book can solve problems related to two or more numbers in LCC, so a unique number assigned to a book is difficult.

VII. SUMMARY

There are also a number of decision making and analysis applications that should be explored in large data sets. For example, in the literature search, we have to find a set of related documents from the millions of document dimensions over 1000. If you get any useful abstraction of the data used by the decision, you can deal with these problems instead of using the entire data set directly. For data abstraction, we refer to a simple and compact representation of the data. This simplicity facilitates the efficient processing of machines or the structure of human data that is easily understood. Clustering algorithm is ideal for the realization of data abstraction.

Clustering is the processing of grouping data elements based on similarity measures. Clustering is a subjective process. The same set of data elements are typically allocated in different ways for different applications. This subject makes the grouping process. This is because a single algorithm or method is not suitable for solving each problem grouping. One possible solution is in the form of knowledge that reacts to this subjectivity. Such knowledge is explicitly or implicitly used in one or more stages of a packet. The knowledge-based clustering algorithm makes explicit use of domain knowledge.

The most difficult step is grouping feature extraction or pattern representation. The researcher pattern recognition is convenient to avoid in this step assuming that the pattern representation can be provided as input to the clustering algorithm. When the datasets are small in size, they can be modeled based on the previous user experience with the problem. However, in the case of large data sets, it is very difficult for the user to track the importance of each feature in the group.

One solution is to make as many possible measurements as possible patterns and use them to represent the employer. But it is not possible to directly measure in the pool, since the calculation uses a large set of costs. As a result, they have designed a variety of method selection / feature extraction that can be used to represent a linear or non-linear combination of patterns. Most of the feature extraction / selection schemes proposed are often iterative in nature and cannot be used together in large datasets because of the high computational cost. The second step in the group is a similar calculation.

They use various schemes to calculate the similarity between the two patterns. Use knowledge either explicitly or implicitly. The knowledge clustering algorithm is used to compute similarity based on explicit knowledge of knowledge. However, if the employer does not use the appropriate characteristics of the representation, it is not possible to obtain significant partitioning regardless of the quality and quantity of knowledge used in the calculation of similarity. There are comparisons between qualitative and quantitative characteristics of the mixed model that do not represent a generally accepted scheme.

REFERENCES

[1] Dr. E. Chandra, V. P. Anuradha, " A Survey on Clustering Algorithms for Data in Spatial Database Management Systems", International Journal of Computer Application, vol. 24, pp. 19-26.
[2] B. Rama, P. Jayashree, S. Jiwani, " A Survey on clustering Current status and challenging issues", International Journal of Computer Science and Engineering, vol. 2, pp. 2976-2980.
[3 "Data Mining and Clustering Techniques", DRTC Workshop on Semantic Web, Bangalore, 2003.
[4] J. G. Augustson and J. Minker. An Analysis of Some Graph Theoretical Clustering Tech-niques. Journal of t] I. K. Ravichandra Rao, he ACM, 17:571-588, 1970.
[5] Rui Xu, Donald C. Wunsch II, "Survey of Clustering Algorithms", IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005.
[6] Stephen E. Cross, editor. IEEE Expert: Special issue on Data Mining. IEEE Computer Society, October 1996.
[7] M. Amadasun and R.A. King, \Low-Level Segmentation of Multispectral Images via Ag-glomerative Clustering of Uniform Neighborhoods," Pattern Recognition 21(3):261{268, 1988.
[8] E. Diday. The Symbolic Approach in Clustering. In H. H. Bock, editor, Classi_cation and Related Methods of Data Analysis, North Holland, Amsterdam, 1988.
[9] G. P. Babu, M. N. Murty, and S. S. Keerthi. Stochastic Connectionist Approach for Pattern Clustering. to appear in IEEE Trans. Systems, Man and Cybernetics, 1997.
[10] E. Backer. Computer-Assisted Reasoning in Cluster Analysis, Prentice Hall, London, 1995.
[11] R. C. Dubes and A. K. Jain. Clustering Methodology in Exploratory Data Analysis. In M. C. Yovits, editor, Advances in Computers, pp. 113{225, Academic Press, New York, 1980.
[12] J. L. Bentley and J. H. Friedman. Fast Algorithms for Constructing Minimal Spanning Trees in Coordinate Spaces. IEEE Trans. Computers, 27:97-105, 1978.
[13] J. C. Bezdek. Pattern Recognition With Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[14] F. B. Backer and L. J. Hubert. A Graph-Theoretic Approach to Goodness-Of-Fit in Complete-Link Hierarchical Clustering. Journal of the American Statistical Association, 71:870-878, 1976.
[15] F. Can. Incremental Clustering for Dynamic Information Processing. ACM Trans. Informa-tion Systems, 11:143-164, 1993.