IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

A Survey on Clustering Methods in Data Mining Techniques

M. Latha[1], Dr. K. Subramaninan[2]

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

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

Abstract: Clustering is a division of data into groups of similar objects. The small number of data represented by the cluster number must be missing some details, but the implementation is simplified. Data model for their cluster. From a practical point of view, clustering plays an important role in data mining applications for scientific data, information retrieval and text mining applications, spatial databases, Web analytics, customer relationship management, marketing, diagnostics and explores medical, computational biology and Many others. Clustering is the subject of active research in various fields such as statistics, pattern recognition and machine learning. The focus of this survey is on data mining clustering. Recently, there have been a variety of algorithms that meet these requirements and have been successfully applied to practical problems in data mining. They are under investigation.

Keywords: Clustering, Data mining, Machine learning, Pattern recognition.

I. INTRODUCTION

The purpose of this survey is to provide a comprehensive review of the different clustering techniques available for data mining. Clustering is a division of data into groups of similar objects. Each group, called the cluster, consists of objects that are similar and dissimilar objects in other groups. The data represent a certain loss of the group with less detail (similar to lossy data compression) but managed to be simplified. It is represented by several sets of data objects, and therefore their data model is conglomerate with many data objects. Clustering data modeling was proposed to be rooted in mathematical, statistical and numerical [1] analysis of historical perspectives. From the point of view of learning machine clusters corresponding to hidden patterns, finding clusters is an unsupervised learning that results from the concept of system data representation. As a result, the group is unsupervised to learn the concept of hidden data. Large databases impose requirements analysis on additional computing cluster data mining transactions. These challenges led to the emergence of a powerful data mining group approach, widely applicable, discussed below.

II. TYPES OF CLUSTERING ALGORITHM

The classification of clustering algorithm is not direct, nor standardized. In fact, the groups overlap again. For the convenience of readers, we provide a classification followed by this survey. The corresponding terms are explained below.

Clustering Algorithms

- Hierarchical Methods
  - Agglomerative Algorithms
  - Divisive Algorithms
- Partitioning Methods
  - Relocation Algorithms
  - Probabilistic Clustering
  - K-Medoids Methods
  - K-Means Methods
  - Density Based Algorithms
- Grid-Based Methods
- Scalable Clustering Algorithms
- Algorithms for High Dimensional Data
  - Subspace Clustering
  - Projection Techniques
  - Co-Clustering Techniques

A. Hierarchical Clustering

Hierarchical clustering creates hierarchies of clusters, or, in other words, a cluster tree, also known as a tree. Each cluster node contains subsets; the sibling clusters are divided by their common father [2] coverage points. This approach allows us to study data at different granularity levels. The advantages of hierarchical clustering include:

• For incorporation into granularity flexibility levels
• Handle any form of similarity or distance easily

The disadvantages of hierarchical clustering involve:

• Standard ambiguity for termination
• Most hierarchical algorithms no longer review the construction (intermediate) facts and their improvements

In hierarchical clustering, our regular point data indicates that the attribute is sometimes secondary. Instead, hierarchical clustering often involves the distance between the training points (dissimilarity) or a similar N × N matrix. Sometimes it is called the connection matrix [3]. The construction of the link metric (see below) The elements of the matrix. It is not realistic to keep such a large array in memory.

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

B. Partitioning Clustering

In this section, we examine data partitioning algorithms into several subsets of data. Since checking the system for all possible subsets is computationally infeasible, some greedy heuristics are used in the iterative optimization form. Specifically, this means that different scenarios move iteratively to reallocate points between clusters [4] K. Different from the traditional hierarchical method, in which the cluster is not built after the review, the demolition algorithm cluster gradually improved. With the proper data, this gives a high quality cluster of results.

One approach to data partitioning is to use a conceptual point of view to identify the unknown parameters that have to be found in a certain model cluster. More specifically, the probability model assumes that the data come from the distribution of different populations and the a priori desired mixture. The corresponding algorithm is described in terms of probabilistic clustering. Probabilistic methods have the obvious advantage of being built on the cluster's interpretability.

a. K-Means Methods

The K-means algorithm is by far the most common clustering tool in scientific and industrial applications. The name comes from the average (or weighted average) points, the so-called centroid CJ for each cluster of CJ. While this obviously does not work well for classifying attributes, it has good geometric and statistical significance for numeric attributes. And the sum of the differences between the centroids is represented by an appropriate distance as the objective function [5]. The sum of the squares of the errors between the points and the corresponding centroids is equal to the total group variance.

The popularity of the K-means algorithm is well deserved. It is simple, straightforward and based on a solid foundation of analysis of variance. The K-means algorithm is also affected from all common suspects:

• The outcome depends largely on the initial guess (or allocation)
• It is known that the calculated local optima is far from complete
• Only numeric attributes are overwritten
• The algorithm lacks scalability

b. K-Medoids Methods

The k medoids method is a group represented by one of its points. We have already mentioned that this is a simple solution because it covers any type of property, and has resistance to temióides that are embedded against the outliers, as peripheral points of the cluster do not affect them. When miedoides are selected, these groups are defined as subsets of the respective madridos near points, and the objective function is defined as the average distance between a point that is not similar to medoid or other measure [6].

PAM is an iterative optimization of the transition between a junction point and a cluster re-nominated as a potential peyideides point of perspective. The guiding principle of the method is the effect of an objective function, which is obviously an expensive strategy [7]. Santa Clara uses several (five) samples, each 40 + 2k points, each of which is carried out by PAM. The data set is assigned to the resulting Madridas, and the objective function is computed and saved as the best madróides system.

c. Density-Based Partitioning

One group is opened in Euclidean space and can be divided into a set of connected components. The idea of a finite set of needs for density, connectivity, and the concept of boundary points is achieved by partitioning. They are closely related to a point nearest neighbor. Defined as a dense connection to component A, the leading density in different directions is increasing. Therefore, density-based algorithms can find clusters of arbitrary shape [8]. This also provides natural protection against extreme values. Figure 4 illustrates the group relocation (e.g., k-means) that produces the problem for partitioning, but some form of clustering is handled correctly by the density-based algorithm. They also have good scalability. These outstanding characteristics and honed certain shortcomings.

C. Grid-Based Methods

In terms of density, connectivity and boundaries, the key concepts in the previous section need to be carefully defined, and they are used. Another way to deal with them is to inherit the underlying spatial properties of the topology. The combination of restricted searches is considered to be a multi-rectangular segment. Keep in mind that a segment (also a cube-cell region). It is a direct Cartesian product of the range of properties (in the case of numerical properties, continuous). For some grading properties commonly used for numerical values, the method partition space is commonly referred to as a grid-based approach [9]. A basic section corresponding to the sub-range in a single pallet or a single value is called a cell.

In general, we are turning our attention to the division of data space. The data partitioning is a segment induction which is generated by the members and thereby separates the space based on the feature separation space of the mesh accumulated from the input data. One advantage of this is that the indirect management of grid data accumulation makes grid-based clustering techniques independent of the ordering of the data. Conversely, the way demolition is done and all incremental algorithms are very sensitive to the sort of data [10]. Although partitioning based on the density function of the numerical attribute is better based on the method of work with different properties of the grid.

D. Co-Occurrence of Categorical Data

In this section we are talking about the concept of classification data, which often involves the fact that resizable transactions are elements of a finite set of elements called universal items. For example, the data

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

from the basket has this form. Each transaction can be presented at a format attribute point, listing all elements JY associated with the transaction binary attribute indicating whether the element j belongs to one or the transaction. This means that it is sparse and the two random transactions have very little in common.

This is why the similarity between them (sub-section approach) is usually determined by the Jekard coefficient. Common examples of these and other types of categorical data Floating-point format attributes are high-dimensional, zero-valued quantities, a few common values between two objects [10] [11]. The classical clustering method does not work well based on similar measures. Since classification / transactional data is important in developing customer profiles, classification planning, analysis and other Web applications, they have developed different clustering approaches based on the idea of co-investment of classified data.

E. Other Clustering Techniques

They have developed a number of other clustering algorithms. Some of the specific requirements for handling applications. Restricted clusters belong to this category. Others are of theoretical interest and are primarily used in other applications that do not belong to data mining. We briefly present these developments in order to monitor learning, gradient descent and the relationship between artificial neural networks and evolutionary methods in the section. Finally, in the development of other sections of the dynamic mentioned is very simple, but not in our ranking match.

a. Constraint-Based Clustering

In a practical application, the customer is the solution, with little limited interest. Clusters are often subject to some specific limitations that make them suitable for the behavioral problems of certain enterprises. . The taxonomy also includes individual group restrictions, which can be described in the aggregate function constraints (minimum, average, etc.) for each group [12]. These restrictions are necessary because they require a new method. In particular, there is a limit to the number of objects counted from below for certain subsets of each cluster (i.e., frequent customers). Cluster partitioning uses iterative optimization to be based on moving objects representing the closest cluster.

b. Relation to Supervised Learning

Both the Forgy algorithm as the EM, the K-means implementation of the iterative optimization. Both models initialize K and take a series of two steps: (1) redistributing the data points (hard or soft), and (2) updating the combined model. The method can be extended to connect the framework with the predicted group. The updated model is considered predictive training based on the target value of the attribute value that is monitored for the current assignment of a classifier. The redistribution point corresponds to the classification prediction using the new training.

c. Gradient Descent and Artificial Neural Networks

If the target k-means clustering function is slightly modified to combine (similar to MS) "fuzzing", ie, if it indicates that the distance is not only used for the nearest neighbor, but also for the centroid,

$$E'(C) = \text{À} \text{À}_{i=1:N} \quad {}_{j=1:k} \|x_i - c_j\|^2 \omega_{ij}^2$$

The exponential À probability Gaussian model is defined on the basis of. This allows the objective function to be microphased relative to the medium and allows the gradient descent method to be applied in general. The vector quantization of the problem is described in detail. The K-means gradient decent approach is called LKMA (local K-means algorithm).

F. Scalability and VLDB Extensions

Clustering algorithms face both scalability problems in terms of computing time and memory requirements. In data mining, reasonable execution time and the ability to use some limited core memory is particularly important. There have been many interesting attempts to extend the subgroup to very large data bases (VLDBs) and can be divided into:

• Incremental mining,
• Flattened data,
• Reliable sampling.

The DIGNET algorithm (with the clustering algorithm "leader" is an incremental supervised learning [13]), which means managing a data point at a time and then discarding an example DIGNET uses k-means to denote iterations without optimization of the centroid. Cluster push or pull depends on whether to loosen or win every point in llegada. Como online clustering only requires a pass through data but largely depends on the sorting of the data and the sub-mass can be caused by the cluster. Can be dynamically born or discarded, and is summarized in the training process, which makes VLDBdinámico. Algunas an additional tool that can be used to increase the yield of clusters it is very attractive.

Pretreatments such as birch are generally based on vector space operations. Also, in many applications, an object (for example, a string) is a metric space. In other words, what we can do with the data points is to calculate the distance between them. Birch's Meimei Bubble metric space presents the VLDB data type. The foliage of each tree is characterized by:

• Points
• Medoid (called clustroid), which provides a minimal error - the square distance between itself and all other points belongs to the leaf
• The radio corresponds to the square root of the average error point

Sampling and adoption of a unique unified control to control the new life of the ample data mining community. This verification is based on Hoffding or Chernov's limit,

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

he said, regardless of the distribution of the actual value of a random variable Y, 0 ≤ Y ≤ R, N independent observations of the average is the actual average I

$$\left|\bar{Y} - \tfrac{1}{n}\mathbf{A}_{j=1:n} Y_j\right| \leq \varepsilon$$

with probability 1-į as soon as

$$\varepsilon = \sqrt{R^2 \ln(1/\delta) / 2n}.$$

These limitations are exploited in the CURE and development of clustering algorithms for predictive mining in scalable decision trees. In the case of balanced packets, a statistical estimate of the sample size is provided. Because of its non-parametric nature, these limitations are ubiquitous.

G. Clustering High Dimensional Data

Data mining objects may have hundreds of properties. The large pool of these dimensions presents a great deal of difficulty, more predictable learning and so on. In the decision tree, for example, do not select properties for splitting nodes that are simply irrelevant, and are known to affect any naive Bayesian. In this group, however, the higher dimensions exhibit even problems. First, any similarity to any definition, attribute-independent existence, eliminates the cluster's tendency to any hope. After all, looking for clusters, there is a desperate company. And this may also increase with the probability of low-dimensional data occurrence, the presence and the number of unrelated attributes. The basic exploratory data analysis (feature selection) prior to the step of grouping is the best way to solve the first problem of unrelated attributes.

H. Dimensionality Reduction

Many spatial clustering algorithms rely on the indexing of spatial data (bar data preparation) to facilitate fast searching for nearest neighbors. Therefore, the index can be relative to the size of the curse of the performance impact play a very good proxy. It is known that sizes of indices below 16 used in clustering algorithms work effectively for size d & gt; 20 and performance degrades to the level search order (albeit a significantly higher limit of the newer coverage index) .

I. Subspace Clustering

Some algorithms are better adapted to higher dimensions. For example, the algorithm cactus (co-occurrence categorical data) fits as well as only clusters that are constrained by the perspective of the 2D projection of the cluster.

CLIQUE begins with a unit of definition - a subspace in a rectangular cell. Only the unit, whose density remains above the threshold τ. A bottom-up approach is appropriate for finding such units. First, a one-dimensional unit divides the compartments with equal widths (gates) in the interval. Both the τ and Ĭ parameters are inputs to the algorithm. The recursive treadmill Q-1-dimensional to q-dimensional units include a first dimension Q-2 general (a priori) from the Q-1 binding unit. All subspaces are ordered by their coverage and at least cover subspace pruning.

The algorithm L'ENCLUS (based on ENTOPY clustering) [Cheng et al. 1999] follow the group's footsteps, instead adopting different criteria for subspace selection. The entropy from the consideration of the entropy: A1, ..., entropy H (A1, ..., AQ) is higher than a threshold low AQ is considered to be good for the group covering the subspace of the attribute. A good subspace of any subspace is also nice because

$$H(A_1,\ldots,A_{q-1}) = H(A_1,\ldots,A_q) - H(A_q \mid A_1,\ldots,A_{q-1}) \leq H(A_1,\ldots,A_q) < \omega.$$

The low entropy subspace corresponds to the skewness distribution density unit. L'ENCLUS computational costs are high.

The algorithm ORCLUS (predicting clustering-oriented) uses a similar approach to design the grouping, but uses non-axial parallel subspace of higher dimensional space. If this is the case, any attribute selection is destined. ORCLUS has clusters that are defined as rendezvous points (zoning), in a subspace that truly owns the sum of square errors (energies) that are lower for a similar k-means transparent pattern. More specifically, for x∈C and E = {E1, ...}} (specific C), the protrusions are defined as {X? E1, ..., X?}. The algorithm sets up an optimal subspace K [14], which is much slower than the left-dimensional dimension of k-clusters. If a suggestion is made to choose a good parameter L, uniformity L is a responsibility.

J. Co-Clustering

In the OLAP attribute, a rollup can be thought of as representing a property group. An interesting general idea of producing a property group with the grouping of points itself leads to the concept of copolymer classes. A common cluster is two points and their attributes are grouped together. This approach reverses the fight: increasing the grouping based on its attribute points, attempting to base the item on the attribute group.

III. GENERAL ALGORITHM ISSUES

We have come up with many different clustering techniques. However, there are some common problems with clustering algorithms that are successfully addressed. Some are ubiquitous, and they are not even specific to unsupervised learning and can be considered part of the overall framework for data mining. Others present some algorithmic solution. VLDB scalability for clustering and high dimensions has been discussed above, but other important issues are discussed below:

- Evaluation of results
- Select the appropriate number of clusters
- data preparation

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 3, Issue 11, November 2016

• Measure approaching
• Handling outliers

A. Assessment of Results
The process of grouping data mining assesses whether any clustering tendencies have a place, all the start, including proper selection of attributes accordingly, and in many cases, the construction of features [15]. With the validation and evaluation system leading to grouping ends. The clustering system can be evaluated by an expert, or by a specific automated procedure. Traditionally, the first type of evaluation involves two issues: (1) cluster interpretation, (2) cluster display. Explanatory depends on the technology used.

B. How Many Clusters?
In many methods the number of cluster ks generated is a user's parameter input. Running an algorithm several times leads to a sequence of cluster systems. Each system consists of finer, less isolated groups. In the case of the k-means, the objective function is monotonically decreasing. Therefore, the answer to this question is what is preferable to the system is not trivial.

C. Data Preparation
The irrelevant attribute makes a successful grouping of useless opportunities because of the negative effects of nearby measures and the elimination of clustering tendencies. Therefore, exploratory data analysis of the sound (EDA) is essential. You can find the overall framework of EDA. As the first order of the enterprise, the EDA eliminates the inappropriate reduction of the retained base and categorical attributes.

D. Handling Outliers
In the amount of noise associated with the application that obtains their measured values of the data, which can be considered as outliers. Alternatively, an outlier may be considered as a legitimate record having an abnormal behavior. In general, clustering techniques do not distinguish between either, either noise or anomaly for grouping. Correspondingly, the preferred method of dealing with outlier data partitioning is to have an extra set of outliers that do not pollute the actual cluster.

IV. CONCLUSION

The overall goal of the data mining process is to separate the information from later use of a large comprehensible data format. The group is an important task in data analysis and data mining. Clustering is the task of a group of objects that are more similar to each other than groups of the same group (clusters). Grouping can use different algorithms, such as hierarchical, partition, network, and algorithmic density. Grouping is a hierarchical clustering-based connection. Hierarchical clustering connection. The partition is based on the center of gravity of the cluster; the value of the K-media set. Density-based groups are defined as high density, and then the rest of the data set is in the area. The algorithm is based on partitioned cluster centroids. The density-based group is defined as the high density, and then the rest of the data is set where it is in the region. Based on the grouping of the grid, the spatial segmentation is formed into a finite number of operations, in all cluster operations, and graphically on the basis of the graphical method of grouping the edge structures based on a set of structural vertex cells, the iterative clustering being generated Techniques for efficient graph clustering of clusters with low cost compared to other clustering techniques.

REFERENCES

[1] ERTOZ, L., STEINBACH, M., and KUMAR, V. 2002. Finding clusters of different sizes, shapes, and densities in noisy, high dimensional data, Technical Report.
[2] Sunith Bandaru Sunith Bandaru, Amos H.C. Ng Kalyanmoy Deb by Data mining methods for knowledge discovery in multi-objective optimization: Part A – Survey,2016.
[3] Assif Assad, Kusum Deep by Applications of Harmony Search Algorithm in Data Mining: A Survey, Springer, 2016.
[4] Emrah Hancer, Dervis Karaboga by A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number, 2016
[5] N. P. Nethravathi, Vaibhav J. Desai, P. Deepa Shenoy, M. Indiramma, K.R. Venugopal by A Brief Survey on Privacy Preserving Data Mining Techniques,2016.
[6] HAREL, D. and KOREN, Y. 2001. clustering spatial data using random walks, InProceedings of the 7th ACM SIGKDD, 281-286. San Francisco, CA.
[7] KOHONEN, T. 2001. Self-Organizing Maps. Springer Series in Information Sciences, 30, Springer.
[8] MANILLA, H. and RUSAKOV, D. 2001. Decomposition of event sequences into independent components. In Proceedings of the 1st SIAM ICDM, Chicago, IL.
[9] MCCALLUM, A., NIGAM, K., and UNGAR, L.H. 2000. Efficient clustering of high-dimensional data sets with application to reference matching.
[10] HAN, J. and KAMBER, M. 2001. Data Mining. Morgan Kaufmann Publishers.
[11] HAN, J., KAMBER, M., and TUNG, A. K. H. 2001. Spatial clustering methods in data mining: A survey. In Miller, H. and Han, J. (Eds.) Geographic Data Mining and Knowledge Discovery, Taylor and Francis.
[12] HAREL, D. and KOREN, Y. 2001. Clustering spatial data using random walks, In Proceedings of the 7th ACM SIGKDD, 281-286. San Francisco, CA.
[13] HEER, J. and CHI, E. 2001. Identification of Webuser traffic composition using multi-modal clustering and information scent. 1st SIAM ICDM, Workshop on Web Mining, 51-58, Chicago, IL.
[14] JEBARA, T. and JAAKKOLA, T. 2000. Feature selection and dualities in maximum entropy discrimination. In Proceedings of the 16th UIA Conference, Stanford, CA.
[15] BRADLEY, P. S., BENNETT, K. P., and DEMIRIZ, A. 2000. Constrained k-means clustering. Technical Report MSR-TR-2000-65. Microsoft Research, Redmond, WA.