A Review on Clustering Technique

Vivek Kumar Sharma, Assistant Professor
Dept. of Computer Science Engineering & Information Technology, Arya College of Engg. And I.T,
Jaipur, India
e-mail: vkshere4every1@gmail.com

Nisha Vasudeva, Assistant Professor
Dept. of Computer Science Engineering & Information Technology, Arya College of Engg. And I.T,
Jaipur, India
e-mail: vasudeva.nisha1@gmail.com

Abstract**—** Hidden Knowledge is very important in data mining field. Large data set have many hidden pattern which have very crucial information, Clustering is such technique which find the hidden pattern from the large data. Artificial Neural Network is very powerful tool in machine learning or in the field of computer visions. Competitive learning is used for Clustering in Neural network. Example of Competitive learning, SOM and ART are famous for clustering. SOM have the limitation of dimension, ART is good but computation cost is very high.

Keywords*-component; clustering, SOM, Neural Network.*

*****

I. Introduction

Humans have been extracting knowledge from data from last decays, but the increasing volume of data in modern times created problem in extracting the knowledge from data. Information leads to power and success, and many technologies are used such as computers, satellites, etc., we have been collecting [2] tremendous amounts of information from these technologies. Initially, with the advent of computers and means for mass digital storage, we started collecting and storing all sorts of data, counting on the power of computers to help sort through this information of data or knowledge of data. Unfortunately, these massive collections of data stored on disparate structures very rapidly became overwhelming. This initial chaos has led to the creation of structured databases and database management systems (DBMS). We have been collecting a myriad of data, from simple numerical measurements and text documents, to more complex information such as spatial data, multimedia channels, and hypertext documents. Non-exclusive list of a variety of information collected in digital form in databases and in flat files business transactions, scientific data, medical and personal data, surveillance video ,pictures, satellite sensing:, games, digital media, CAD and software engineering [3] data, virtual worlds, text reports and memos (e-mail messages), World Wide Web repositories.

Early methods of identifying patterns in data include bayesian theorem and regression analysis . The proliferation, ubiquity and increasing power of computer technology has increased data collection and storage. As data sets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering, genetic algorithms (1950s), decision trees (1960s) and support vector machines (1980s). Data mining is the process of applying these methods to data with the intention of uncovering hidden patterns.[1,2] With the amount of data stored in files, databases, and other repositories, it is increasingly important, if not necessary, to develop powerful means for analysis and perhaps interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially[3] useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.

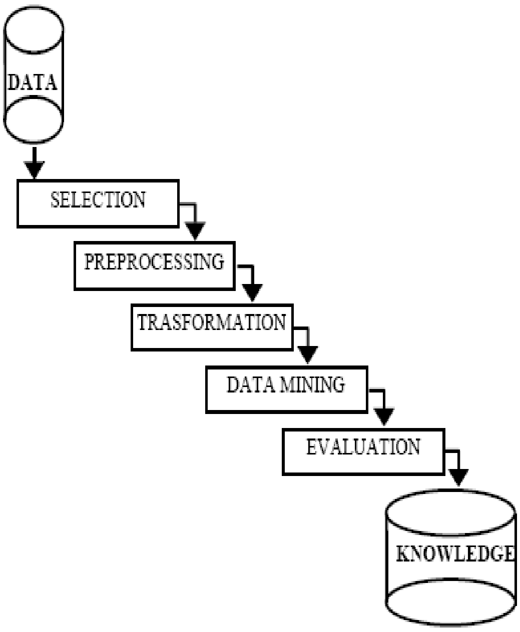

Fig. 1: The Knowledge Discovery Process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

- Data cleaning: also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

- Data integration: in this step, multiple data sources, often heterogeneous, may be combined in a common source.
- Data selection: at this step, the data relevant to the analysis is decided on and retrieved from the data collection.
- Data transformation: also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
- Data mining: it is the crucial step in which clever techniques are applied to extract patterns potentially useful.
- Pattern evaluation: in this step, strictly interesting patterns representing knowledge are identified based on given measures.
- Knowledge representation: is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

Data mining derives its name from the similarities between searching for valuable information in a large database. Both imply either sifting through a large amount of material or ingeniously probing the material to exactly pinpoint where the values reside. It is, however, a misnomer, since mining for gold in rocks is usually called "gold mining" and not "rock mining", thus by analogy, data mining should have been called "knowledge mining" instead. The kinds of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: descriptive data mining tasks that describe the general properties [3] of the existing data, and predictive data mining tasks that attempt to do predictions based on inference on available data.

II. CLUSTERING IN DATA MI NING

Clustering in data mining group similar objects into clusters. A cluster is therefore a collection of objects which are similar between them and are dissimilar with objects belonging to other clusters. Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification. From a machine learning perspective clusters correspond to hidden [2] patterns, the search for clusters is unsupervised learning, and the resulting system represents a data concept. From a practical perspective clustering plays an outstanding role in data mining applications such as scientific data exploration, information retrieval and text mining, spatial database applications, Web analysis, CRM, marketing, medical diagnostics, computational biology, and many others.

Unlike classification, clustering is a type of unsupervised learning. It groups similar data items into clusters without knowing their class membership. The basic principle is to maximize intra-cluster similarity while minimizing inter-cluster similarity. Clustering has been used in a variety of applications, including image segmentation gene clustering and document [3] categorization. Various clustering methods have been developed, including *hierarchical approaches* such as complete-link algorithms, *partitional approaches* , and *Self-Organizing Maps*. These clustering methods group data items based on different criteria and may not generate the same clustering results. Hierarchical clustering groups data items into a series of nested clusters and generates a tree-like dendrogram.

Partitioned clustering algorithms generate only one partition level rather than nested clusters. Partitioned clustering is more efficient and scalable for large datasets than hierarchical clustering, but has the problem of determining the appropriate number of clusters. Different from the hierarchical and Partitioned clustering that relies on the similarity or proximity measures between data items, SOM is a neural network-based approach that directly projects multivariate data [2] items onto two-dimensional maps. SOM can be used for clustering and visualizing data items and groups. SOM is popular in vector quantization. SOM has two important features: (1) SOM uses incremental approach. points (patterns) are processed one-by-one; (2) SOM allows mapping centroids into 2D plane that provides for a straightforward visualization. In addition to SOM, other ANN developments, such as adaptive resonance theory [Carpenter et al. 1991], have relation to clustering. *K*means [2],fuzzy cmeans [3, 4], Self Organizing Map (SOM) [5, 6], Artificial Neural Networks [7] and Support Vector Machines [8]. These techniques belong to one of two groups supervised clustering and unsupervised clustering depending upon the underlying method used to cluster the data items. In supervised clustering, the algorithm is trained using a proportion of the dataset (the training set) and then this trained algorithm is used to classify an unknown dataset (test set). Typically, the number of clusters for supervised learning is pre-specified. Alternatively, in unsupervised clustering a training dataset is not used**.** The accuracy and the generalization capability of SOM are solely dependent on the topology of the map chosen and the map size.

III. RELATED WORK

A. Clustering in Neural Network

Artificial neural networks (ANNs) are motivated by biological neural networks. ANNs have been used extensively over the past three decades for both classification and clustering. As an unsupervised classification technique, clustering identifies some inherent structures present in a set of objects based on a similarity measure.[1] Most datasets in real applications come in from multiple sources. As a result, we often have attributes information about data objects and various pair wise relations (similarity) between data objects. Traditional clustering algorithms use either data attributes only or pair wise similarity [2].Large datasets can be analyzed through different linear and nonlinear methods.

B. Principal Component Analysis

Principal Component Analysis (PCA) also known as EOF (Empirical Orthogonal Function) analysis, permitting both clustering and visualizing large data items. However, many problems are nonlinear in nature, so, for analyzing such problems some nonlinear methods will be more appropriate [3]. Principal Components Analysis is used to reduce data dimensionality by performing a covariance analysis between factors. As such, it is suitable for data sets in multiple dimensions, such as a large experiment in gene expression [4] the application of sparse principal component analysis (PCA) to clustering and feature selection problems. Sparse PCA seeks sparse factors, or linear combinations of the data variables, explaining a maximum amount of variance in the data while having only a limited number of nonzero coefficients. PCA is often used as a simple clustering technique and sparse factors allow us here to interpret the clusters in terms of a reduced set of variables [5].

C. Kohonen Network

In 1973, by von der Malsburg, is topology-preserving competitive learning models that are inspired by the cortex of mammals. The Kohonen network has the same structure as the competitive learning network. The output layer is called the Kohonen layer. Lateral connections are used as a form of feedback whose magnitude is dependent on the lateral distance from a specific neuron, which is characterized by a neighborhood parameter.

D. Self organising Map

The Kohonen network is called the SOM when the lateral feedback is more sophisticated than the WTA rule..The SOM (Self-Organizing Map) neural network is very promising tool for clustering and mapping spatial-temporal datasets describing nonlinear phenomena.[6] SOM algorithm is applicable to large data sets However, to be able to fully understand contents of a data set, it is vital to find out if the data has cluster structure. The computational[7,8] complexity scales linearly with the number of data samples, it does not require huge amounts of memory basically just the prototype vectors and the current training vector and can be implemented both in a neural, on-line learning manner as well as parallelized [7,9] . The SOM is designed for real-valued vectorial data analysis, and it is not suitable for non-vectorial data analysis such as the structured data analysis.

E. Support Vector Machine

Support Vector Machines (SVM) classifiers with very large datasets. this is used to preprocess standard training data and simply extended to deal with clustered data, that is effectively a set of weighted examples there is also a Clustering-Based SVM (CB-SVM), which is specifically designed for handling very large data sets. CB-SVM applies a hierarchical micro-clustering algorithm that scans the entire data set only once to provide an SVM with high quality samples that carry the statistical summaries of the data such that the summaries maximize the benefit of learning the SVM. CB-SVM tries to generate the best SVM boundary for very large data sets given limited amount of resources.[18] Traditional clustering algorithms use either data attributes only or pairwise similarity only.

F. Vector Quantization

Vector quantization (VQ) is a classical method for approximating a continuous probability density function (PDF). In 1979 by Gersho, The space is partitioned into a finite number of regions bordered by hyper-planes. Each region is represented by a codebook vector, which is the nearest neighbor to any point within the region. Given a competitive learning based clustering method, learning is first conducted to adjust the algorithmic parameters; after the learning phase is completed, the network is ready for generalization. When a new input pattern x is presented to the map, the map gives the corresponding output c based on the nearest neighborhood rule.

G. Learning vector quantization

In 1973 by Duda and Hart, the k-nearest-neighbor (k-NN) algorithm is a conventional classification technique. It is also used for outlier detection. There are two families of the LVQ- Style models, supervised models such as the LVQ1, th LVQ2, and the LVQ3 as well as unsupervised models such as the LVQ in 1989 by Kohonen and the incremental C-means in 1967 by MacQueen, LVQ2 and LVQ3 have the problem of reference vector divergence

H. ART networks

In 1976 by Grossberg, Adaptive resonance theory (ART) is biologically motivated and is a major advance in the competitive learning paradigm. The ART has the ability to adapt, yet not forget the past training, and it overcomes the so-called stability plasticity dilemma. ART model family includes a series of unsupervised learning models. ART networks employ a J-K recurrent architecture. The input layer F1, called the comparing layer, has J neurons while the output layer F2, called the recognizing layer, has K neurons. F1 and F2 are fully interconnected in both directions. F2 acts as a WTA network. There are various modification and changes are done in ART in the recent years.

I. CLARANS

(Clustering Large Applications based upon Randomized Search) [7] is a partitioning clustering algorithm developed for large data sets, which uses a randomized and bounded search strategy to improve the scalability of the k-medoid approach. CLARANS enables the detection of outliers and its computational complexity is about O(n^2). CLARANS' performance can be improved by exploring spatial data structures such as R*-trees. Hierarchical clustering algorithms work by grouping data objects into a hierarchy (e.g., a tree) of clusters. The hierarchy can be formed top-down (divisive hierarchical methods) or bottom-up (agglomerative hierarchical methods).

J. Single Linkage Algprithm

The single linkage clustering method is the simplest of all hierarchical agglomerative methods, also known as nearest neighbor technique first described by Florek et al . The defining feature of the method is the distance between two clusters defined as the distance between the closest data elements of the two clusters and so the rest of the data elements of the clusters has nothing in the calculation of the inter cluster separation.

K. Complete Linkage algorithm

The complete linkage clustering methods is also called the furthest neighbor clustering method. It is a hierarchical agglomerative method where the distance between two clusters to be merged is calculated using the distance between the two farthest data elements of the two clusters. It is an exact opposite strategy to that of the single linkage clustering method.

Hierarchical methods rely on a distance function to measure the similarity between clusters. These methods do not scale well with the number of data objects. Their computational complexity is usually O(n^2). Some newer methods such as BIRCH and CURE attempt to address the scalability problem and improve the quality of clustering results for hierarchical methods. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) is an efficient divisive hierarchical algorithm. It has O(n) computational complexity, can work with limited amount of memory, and has efficient I/O. It uses a special data structure, CF tree (Cluster Feature tree) for storing summary information about sub clusters of objects.
The CF-tree structure can be seen as a multilevel compression of the data that attempts to preserve the clustering structure inherent in the data set. Because of the similarity measure it uses to determine the data items to be compressed, BIRCH only performs well on data sets with spherical clusters. CURE (Clustering Using RE presentatives) is an O(n^2) algorithm that produces high-quality clusters in the presence of outliers, and can identify clusters of complex shapes and different sizes. It employs a hierarchical clustering approach that uses a fixed number of representative points to define a cluster instead of a single centroid or object. CURE handles large data sets through a combination of random sampling and partitioning. Since CURE uses only a random sample of the data set, it manages to achieve good scalability for large data sets. CURE reports better times than BIRCH on the same benchmark data.

Locality-based clustering algorithms group neighboring data objects into clusters based on local conditions. These algorithms allow clustering to be performed in one scan of the data set. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) [10] is a typical representative of this group of algorithms. It regards clusters as dense regions of objects in the input space that are separated by regions of low density. DBSCAN's basic idea is that the density of points in a radius around each point in a cluster has to be above a certain threshold. It grows a cluster as long as, for each data point within this cluster, a neighborhood of a given radius contains at least a minimum number of points. DBSCAN has computational complexity O(n^2). If a spatial index is used, the computational complexity is O(n log n). The clustering generated by DBSCAN is very sensitive to parameter choice. OPTICS (Ordering Points to Identify Clustering) is proposed to compute a density-based hierarchical decomposition of the data

In general, partitioning, hierarchical, and locality-based clustering algorithms do not scale well with the number of objects in the data set. To improve the efficiency, data summarization techniques integrated with the clustering process have been proposed. Besides the above-mentioned BIRCH and CURE algorithms, examples include: active data clustering [5], Scalable, and simple single pass k-means. Active data clustering utilizes principles from sequential experimental design in order to interleave data generation and data analysis. It infers from the available data not only the grouping structure in the data, but also which data are most relevant for the clustering problem. The inferred relevance of the data is then used to control the re-sampling of the data set. Scalable-KM requires at most one scan of the data set.

The method identifies data points that can be effectively compressed, data points that must be maintained in memory, and data points that can be discarded. The algorithm operates within the confines of a limited memory buffer. Unfortunately, the compression schemes used by Scalable-KM can introduce significant overhead. The simple single pass k-means algorithm is a simplification of Scalable-KM. Like Scalable-KM, it also uses a data buffer of fixed size. Experiments indicate that the simple single pass k-means algorithm is several times faster than standard k-means while producing clustering of comparable quality.

In this regard, some attempts have been made to use genetic algorithms for automatically clustering data sets [2]. Genetic algorithms (GA's) work on a coding of the parameter set over which the search has to be performed, rather than the parameters themselves [1]. These encoded parameters are called solutions or chromosomes and the objective function value at a solution is the objective function value at the corresponding parameters. GA's solve optimization problems using a population of a fixed number, called the population size, of solutions. A solution consists of a string of symbols, typically binary symbols. GA's evolve over generations. During each generation, they produce a new population from the current population by applying genetic operator's viz., natural selection, crossover, and mutation .

Each solution in the population is associated with a figure of merit (fitness value) depending on the value of the function to be optimized. The selection operator selects a solution from the current population for the next population with probability proportional to its fitness value. Crossover operates on two solution strings and results in another two strings. Typical crossover operator exchanges the segments of selected strings across a crossover point with a probability. The mutation operator toggles each position in a string with a probability, called the mutation probability.

Bandyopadhyay and Maulik [6] applied the variable string length genetic algorithm with the real encoding of the coordinates of the cluster centers in the chromosome to the clustering problem. A K-mean clustering is performed taking the membership degrees and prototype locations as the parameters for the GA in [8]. In an extension, the prototype locations are encoded as binary strings and genetic operators then operate to optimize location instead of operating on the membership matrices. The prototype location is also encoded in as real-numbered ordered pairs of length k*d, where k is the number of clusters and d is the dimensionality of the dataset.

IV. CONCLUSION

In this paper, we give a brief introduction to Data mining and knowledge discovery. Brief detail is discussed of clustering techniques. Hierarchical based method, partition based method, model based techniques are basic three types of clustering. Traditional method of clustering is not accurate and efficient for large data sets. Number of iteration increase as the size is increase.

REFERENCES

[1] Y. Cheng and G.M. Church. Biclustering of expression data. In Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology (ISMB), pages 93–103, San Diego, USA, 2000.

[2] Jiawai Han and Micheline Kamber, Data mining-concepts and techniques

[3] WH Inmon,Wiley, Building the Data Warehouse

[4] H.S. Nagesh, S. Goil, and A.N. Choudhary. A scalable parallel subspace clustering algorithm for massive data sets. In International Conference on Parallel Processing, pages 477–, Toronto, Canada, 2000.

[5] K.G.Woo J.H. Lee. Findi:a fast and intelligent subspace clustering algorithm using dimension voting. Information Software Technology, 46:255–271, 2004.

[6] R. Aggarwal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining application, In Proc. of ACM SIGMOD Conference, 1998.

[7] Eric Ka Ka Ng, Ada Wai-chee Fu and Raymond Chi-Wing Wong, "Projective clustering by histograms", IEEE Transactions on Knowledge and Data Engineering, Volume: 17, Issue: 3, pp 369-383, March 2005

[8] C.Aggarwal andP.Yu. "Finding Generalized Projected Clusters in High Dimensional Space". In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD' 00), 2000.

[9] C. C. Aggarwal and C. Procopiuc. "Fast Algorithms for Projected Clustering". In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'99), 1999.

[10] R.Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications". In Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD' 98), 1998.

[11] Neural Networks, IEEE Transactions on On ,Clustering of the self-organizing map Vesanto,JAlhoniemi,E, Neural Networks Res. Centre, Helsinki Univ. of Technol. page(s): 586-600 Volume: 11, May 2000 ISSN: 1045-9227

[12] Geoscience and Remote Sensing Symposium, 2000. Proceedings. IGARSS 2000. IEEE 2000 International 07/24/2000, DGLC: a density-based global logical combinatorial clustering algorithm for large mixed incomplete Ruiz-Shulcloper, J. Alba-Cabrera, E. Sanchez-Diaz,G. Dept. of Electr. & Comput. Eng., Tennessee Univ., Knoxville 2000 Location: Honolulu, HI , USA On page(s): 2846-2848 vol.7

Bibliography

Nisha Vasudeva received the M.Sc (Computer Science) from Maharshi Dayanand University in 2008 and M.Tech degree in Computer Science Engineering from Mody Institute of Science and Technology, Laxmangarh in 2011.I am currently working in Arya college of Engg. And Tech., Jaipur, Rajasthan. I have experience of 4 years.

Vivek Kumar Sharma received the B.Tech from Rajasthan University in 2008 and M.Tech degree in Computer Science Engineering from JaganNath University , Jaipur in 2012.I am currently working in Arya college of Engg. And Tech., Jaipur, Rajasthan. I have experience of 7 years in teaching.