Assessment of Clustering Techniques in Data Mining

RAMYA M.C, DEBABRATA SAMANTA
Department of MCA, Acharya Institute of Technology (AIT), Bangalore, India

ABSTRACT

Cluster study or clustering is the assignment of assigning a set of data into groups called clusters. Major task of clustering are explorative data mining, and a frequent methodology for statistical data analysis used in different fields, pattern recognition, image analysis, including machine learning, information retrieval, and bioinformatics. Cluster can be accomplished by diverse algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters embrace groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. There are numerous no of classification mechanisms that can be used such as K-nearest neighbor, Bayesian network, Neural Networks, Decision Trees, Support Vector Machine, and Boosting etc. This paper also compacts with a study on some of these commonly used techniques that are being widely used.

KEYWORDS: *Decision Tree, Data Mining, Clustering Techniques, Performance Analysis, Bayesian network, K-nearest neighbor*

I. INTRODUCTION

A process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing. In addition to collecting and managing of data, data mining also includes analysis and prediction.
Classification techniques in data mining are talented of processing a bulky amount of data. It can envisage categorical class labels and classifies data based on training sample of data and class labels and hence can be used for classifying newly existing data. Thus it can be outlined as an inevitable part of data mining and is gaining more popularity. In the present paper a study on various classification techniques have been made. Next section deals with some comparisons.

II. BACKGROUND CLUSTERING TECHNIQUES

2.1. Hierarchical clustering

Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The basic process of hierarchical clustering:

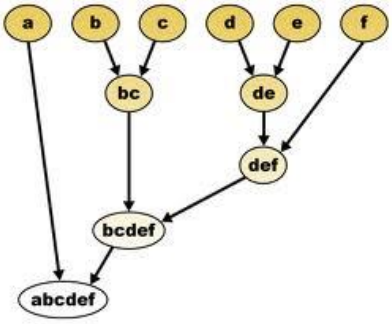
1. Start by assigning each item to a cluster, so that if you have N items, you now have N clusters, each containing just one item. Let the distances (similarities) between the clusters the same as the distances (similarities) between the items they contain.
2. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one cluster less.
3. Compute distances (similarities) between the new cluster and each of the old clusters.

Repeat steps 2 and 3 until all items are clustered into a single cluster of size N. (*)
Strategies for hierarchical clustering generally fall into two types. The results of hierarchical clustering are usually presented in a dendrogram. Hierarchical clustering does not require us to pre specify the number of clusters and most hierarchical algorithms that have been used in IR are deterministic. Hierarchical clustering method is useful to cluster categorical or mixed data. However, the hierarchical clustering method is not efficient in processing large data sets. Their use is limited to

small data sets result of the hierarchical methods is a dendrogram, representing the nested grouping of objects and similarity levels at which groupings change.

2.2. Agglomerative

Agglomerative hierarchical clustering is a bottom-up clustering method where clusters have sub-clusters, which in turn have sub-clusters, etc. The classic example of this is species taxonomy. Gene expression data might also exhibit this hierarchical quality (e.g. neurotransmitter gene families). Agglomerative hierarchical clustering starts with every single object (gene or sample) in a single cluster. Then, in each successive iteration, it agglomerates (merges) the closest pair of clusters by satisfying some similarity criteria, until all of the data is in one cluster.



2.3. Centroid-based clustering

The centroid of a cluster is a point whose parameter values are the mean of the parameter values of all the points in the clusters. Generally, the distance between two points is taken as a common metric to assess the similarity among the components of a population. The commonly used distance measure is the Euclidean metric which defines the distance between two points p= (p1, p2, ...) and q = (q1, q2, ...) is given by :

$$d = \sqrt{\sum_{r=1}^{n} (P_r - Q_c)^2}$$

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized. The centroid represents the most typical case in a cluster. For example, in a data set of customer ages and incomes, the centroid of each cluster would be a customer of average age and average income in that cluster. If the data set included gender, the centroid would have the gender most frequently represented in the cluster.

The Center-based Clustering algorithm normalizes input variables to the value range [0;1]. Categorical input variables are encoded by using nominal encoding. Therefore categorical input variables with lots of different values can slow down the mining run considerably. With Center-based Clustering, you must specify a total number of passes. With each pass, the center vectors are adjusted to minimize the total distance between cluster centers and records. Also, the amount by which the vectors are adjusted is decreased. The following parameters are specific to Center-based Clustering:

- Map layout
- Number of passes
- Maximum number of clusters

2.4. Distribution-based clustering

Distribution-based Clustering provides fast and natural clustering of very large databases. It automatically determines the number of clusters to be generated. Distribution-based Clustering is an iterative process over the input data. Each input record is read in succession. The similarity of each record with each of the currently existing clusters is calculated. Initially, no clusters exist. If the biggest calculated similarity is above a given threshold, the record is added to the relevant cluster.

This cluster's characteristics change accordingly. If the calculated similarity is not above the threshold, or if there is no cluster, a new cluster is created, containing the record alone. You can specify the maximum number of clusters, as well as the similarity threshold.

Distribution-based Clustering provides fast and natural clustering of very large databases. It automatically determines the number of clusters to be generated. Distribution-based Clustering uses the statistical Condorcet criteria to manage the calculation of the similarity between records and other records, between records and clusters, and between clusters and other clusters. The Condorcet criteria evaluates how homogeneous each discovered cluster is (in that the records it contains are similar) and how heterogeneous the discovered clusters are among each other. The iterative process of discovering clusters stops after one or more passes over the input data if there is no time remaining to do another pass or if the improvement of the clusters according to the Condorcet criteria would not justify a new pass. The following parameters are specific to Distribution-based Clustering:

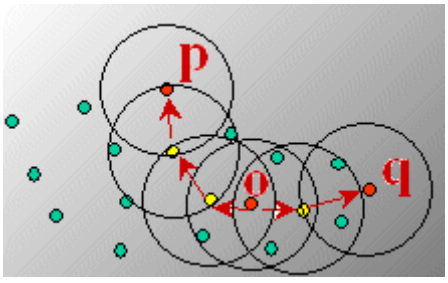
- Similarity threshold
- Similarity scale
- Similarity matrices
- Field weighting
- Value weighting
- Maximum number of clusters

2.5. Density-based clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points. Density-based approaches apply a local cluster criterion. Clusters are regarded as regions in the data space in which the objects are dense, and which are separated by regions of low object density (noise). These regions may have an arbitrary shape and the points inside a region may be arbitrarily distributed.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

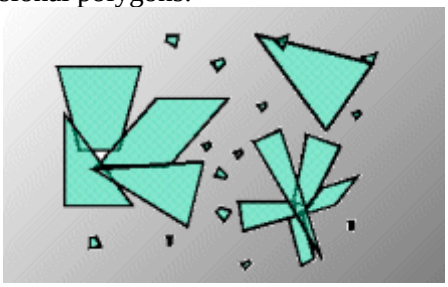
The algorithm DBSCAN, based on the formal notion of density-reach ability for k-dimensional points, is designed to discover clusters of arbitrary shape. The runtime of the algorithm is of the order O(n log n) if region queries are efficiently supported by spatial index structures, i.e. at least in moderately dimensional spaces.



Example of DBSCAN

GDBSCAN (Generalized Density-Based Spatial Clustering of Applications with Noise)

GDBSCAN generalizes the notion of point density and therefore it can be applied to objects of arbitrary data type, e.g. 2-dimensional polygons.



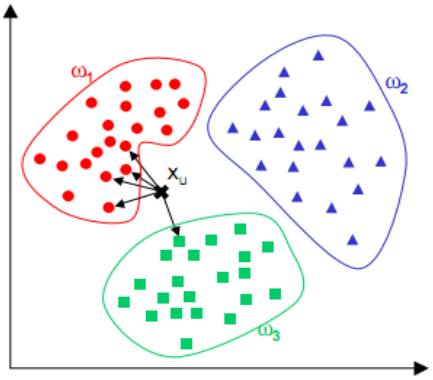
Example of GDBSCAN

III. SURVEY OF CLUSTERING TECHNIQUES

3.1. K-nearest neighbor

Features of K-nearest neighbor are given below.

- All instances correspond to points in an n-dimensional Euclidean space
- Classification is delayed till a new instance arrives
- Classification done by comparing feature vectors of the different points
- Target function may be discrete or real-valued



Example of K-NN classification

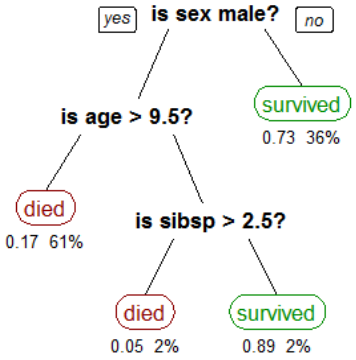
k-means clustering in particular when using heuristics such as Lloyd's algorithm is rather easy to implement and apply even on large data sets. As such, it has been successfully used in various topics, ranging from market segmentation, computer vision, and astronomy to agriculture. It often is used as a preprocessing step for other algorithms, for example to find a starting configuration.

3.2. Mean shift clustering

Basic mean shift clustering algorithms maintain a set of data points the same size as the input data set. Initially, this set is copied from the input set. Then this set is iteratively replaced by the mean of those points in the set that are within a given distance of that point. A mean shift algorithm that is similar then to k-means, called likelihood mean shift, replaces the set of points undergoing replacement by the mean of all points in the input set that are within a given distance of the changing set. One of the advantages of mean shift over k-means is that there is no need to choose the number of clusters, because mean shift is likely to find only a few clusters if indeed only a small number exist. However, mean shift can be much slower than k-means, and still requires selection of a bandwidth parameter. Mean shift has soft variants much as k-means does.

3.3. Decision tree

Decision tree learning is a method commonly used in data mining. The goal is to create a model that predicts the value of a target variable based on several input variables.

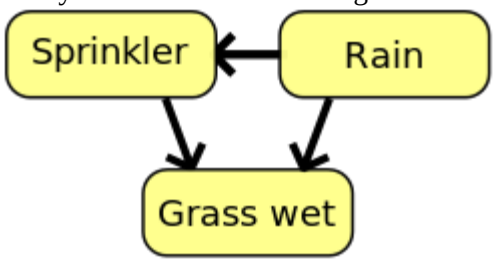


Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. If in practice decisions have to be taken online with no recall under incomplete knowledge, a decision tree should be paralleled by a probability model as a best choice model or online selection model algorithm. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees, influence diagrams, utility functions, and other decision analysis tools and methods are taught to undergraduate students in schools of business, health economics, and public health, and are examples of operations research or management science methods.

3.4. Bayesian network

A Bayesian network, Bayes network, belief network, Bayes(ian) model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Efficient algorithms exist that perform inference and learning in Bayesian networks. Bayesian networks that model sequences of variables (e.g. speech signals or protein sequences) are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.



3.5. Artificial neural network

The term neural network was traditionally used to refer to a network or circuit of biological neurons. The modern usage of the term often refers to artificial neural networks, which are composed of artificial neurons or nodes. Thus the term may refer to either biological neural networks, made up of real biological neurons, or artificial neural networks, for solving artificial intelligence problems. The connections of the biological neuron are modelled as weights. A positive weight reflects an excitatory connection, while negative values mean inhibitory connections. All inputs are modified by a weight and summed altogether. This activity is referred as a linear combination.

IV. COMPARISON OF CLASSIFICATION TECHNIQUES

Method	Generative Or Discriminative	Parameter estimation algorithm	Loss Function
K-Nearest Neighbor	Discriminative	Store all training data to classify new points.	-log P(X,Y) Or Zero-one loss
Bayesian Network	Generative	Variable Elimination	-log P(X,Y)
Decision Tree	Discriminative	C4.5	Zero-one loss
Neural Network	Discriminative	Forward Propagation	Sum-squared error

V. CONCLUSION

This dissertation deals with different classification techniques used in data mining. Each of these techniques can be used in different circumstances as needed where one tends to be useful while the other may not and vice-versa. Hence these classification techniques show how a data can be determined and grouped when a new set of sample data is available. Each methodology has got its

own pros and cons as given in the paper. Based on the needed conditions each one as needed can be selected.

REFERENCES

[1] A Fast Decision Tree Learning Algorithm Jiang Su and Harry Zhang Faculty of Computer Science University of New Brunswick, NB, Canada, E3B 5A3.
[2] K.-L. Tan, P.-K. Eng, and B.C. Ooi, "Efficient Progressive Skyline Computation," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2001.
[4] Charniak, E. 1991, .Bayesian Networks without tears. AI Magazine, Winter 1991.
[5] Ben-Gal I., Bayesian Networks, in Ruggeri F., Faltin F. & Kenett R Encyclopedia of Statistics in Quality & Reliability, Wiley & Sons (2007).
[6] Jordan, M.I. (1999). Learning in Graphical Models, MIT Press, and Cambridge.
[7] Maryam hajiee, "A New Distributed Clustering Algorithm Based on K-means Algorithm", 2010 3rd International Conference on Advanced Computer Theory and Engineering (1CACTE), pp. 408-411 (V2).
[8] Madjid Khalilian, Farsad Zamani Boroujeni, Norwati Mustapha, Md. Nasir Sulaiman, "K-Means Divide and Conquer Clustering", IEEE 2009, International Conference on Computer and Automation Engineering, pp. 306-309.
[9] F. Yang, T. Sun, C. Zhang, An efficient hybrid data clustering method based on K-harmonic means, and Particle Swarm Optimization, Expert Systems with Applications 2009, pp. 9847–9852.
[10] Y.-T. Kao, E. Zahara, I.-W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 2008, pp. 1754-1762.
[11] T.H. Cormen, C.E. Leiserson, R.L. Rivest, and C. Stein, Introduction to Algorithms. MIT Press, 2001.
[12] D. Kossmann, F. Ramsak, and S. Rost, "Shooting Stars in the Sky: An Online Algorithm for Skyline Queries," Proc. Int'l Conf. Very Large Data Bases (VLDB), 2002.
[13] Ajith Abraham and Ravi Jain. Soft Computing Models for Network Intrusion Detection Systems.
[14] Jose F. Nieves (2009), Data Clustering for Anomaly Detection in Network Intrusion Detection.

AUTHORS

Ramya M C. obtained Master degree in Computer science in Mysore University, India in 2006. Major research areas include Data mining. She is working as Assistant professor in Acharya Institute of technology, Bangalore from 2007 to date She has guided many UG projects. She is a Life member of Indian Society for Technical Education. She has 03 National conference publications.

Debabrata Samanta, a member of the IAENG, Board member of the Seventh Sense Research Group Journals (SSRGJ). He obtained my B.Sc. (Physics Honors) in the year 2007, from the Vivekananda Collage, Takurpukur, under Calcutta University; Kolkata, India .He obtained my MCA in the year 2010, from the Academy Of Technology, under WBUT. He has been working his PhD in Computer Science and Engg. from the year 2010 from National Institute of Technology, Durgapur, India in the area of Image Processing .He is presently working as a Assistant Professor Grade III of MCA dept in Acharya Institute of Technology, Bangalore, Karnataka, India from 19th Aug,2013. His areas of interest are Artificial Intelligence, Natural Language Processing and Image Processing. He has published 45 papers in International Journals / Conferences.