International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600

http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-280 PAGE NO: 1819-1829

Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

ANALYSIS OF CLUSTERING AND CLASSIFICATION

L.Visalatchi[1], VR.Deepalakshmi[2],S.Santhi[3]

[1]Associate professor, [2]MSc-InformationTechnology,[3]MSc-InformationTechnology
[1,2,3]Department of Information Technology, Dr.UmayalRamanathan College For Women,
Karaikudi-600 003, Tamilnadu, India.
[1]visaramki@gmail.com,[2]deepaVeerappan96@gmail.com

Abstract- Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were more time consuming to resolve.

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clustercentre that will represent each cluster.

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data.The term could cover any context in which some decision or forecast is made on the basis of presently available information. Classification procedureis recognized method for repeatedly making such decisions in new situations.

Clustering

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster.clustering is the process of making a group of abstract objects into classes of similar objects.

- A cluster of data objects can be treated as one group.
- While doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

- The main advantage of clustering over classification is that, it is adaptable to changes and helps single out useful features that distinguish different groups.

why clustering?

- Scalability
- Ability to deal with different kinds of attributes
- Discovery of clusters with attribute shape
- High dimensionality
- Ability to deal with noisy data
- Interpretability

Clustering Methods

Clustering methods can be classified into the following categories −

- Partitioning Method
- Hierarchical Method
- Density-based Method
- Grid-Based Method
- Model-Based Method
- Constraint-based Method

Partitioning Method

In the partitioning method partitions set of n data objects into k clusters such that all the data objects into same clusters are closer to center mean values so that the sum of squared distance from mean within each clusters is minimum. There are two types of partitioning algorithm.
1) Center based k-mean algorithm
2) Medoid based k-mode algorithm.

- Each group contains at least one object.
- Each object must belong to exactly one group.

Pros:

- It is easy to understand and implement.
- It takes less time to execute as compared to other techniques.
- It can handle large data sets.
- It can handle with only categorical values.

Cons:

- User has to provide pre-determined value of k.
- It produces spherical shaped clusters.
- It cannot handle with noisy data objects.
- The order of data objects have to maintain.

Hierarchical Methods

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed. There are two approaches here −

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach

This approach is also known as the bottom-up approach. In this, we start with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keep on doing so until all of the groups are merged into one or until the termination condition holds.

Divisive Approach

This approach is also known as the top-down approach. In this, we start with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering

Here are the two approaches that are used to improve the quality of hierarchical clustering −

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

here are three measure of cluster proximity**: single-link, complete-link and average-link**

Single-link: The distance between two clusters to be the smallest distance between two points such that one point is in each cluster.
Complete-link: The distance between two clusters to be the largest distance between two points such that one point is in each cluster.
Average-link: The distance between two clusters to be an average distance between two points such that one point is in each cluster.

Pros:

- ➢ It produces clusters of arbitrary shapes.
- ➢ It can handle noise in the data sets effectively.
- ➢ It can handle with outliers.

Cons:

- ➢ The steps during merge and split process cannot be undone.
- ➢ It cannot handle with noisy data objects.

Density-based Method

These algorithms create clusters according to the high density of members of a data set, in a determined location.This method is based on the notion of density. The basic idea is to continue growing the given cluster as long as the density in the neighborhood exceeds some threshold, i.e., for each data point within a given cluster, the radius of a given cluster has to contain at least a minimum number of points.

Pros:

- The number of clusters is not required.
- It can handle large amount of noise in data set.
- It produces arbitrary shaped clusters.
- It is most insensitive to ordering of data objects in dataset.

Cons:

- Quality of clustering depends on distance measure.

Grid-based Method

In this, the objects together form a grid. The object space is quantized into finite number of cells that form a grid structure.The main focus of these algorithms is spatial data, i.e, data that model the geo-metric structure of objects in space, their relationships, properties and operations.In this sense, they are closer to hierarchical algorithms but the merging of grids, and conse-quently clusters, does not depend on a distance measure but it is decided by a predefined parameter

pros:

- The major advantage of this method is fast processing time.
- It is dependent only on the number of cells in each dimension in the quantized space.

Cons:

- This methods depends only the number of cells in each dimension in the quantized space.

Model-based methods

In this method, a model is hypothesized for each cluster to find the best fit of data for a given model. This method locates the clusters by clustering the density function. It reflects spatial distribution of the data points.

This method also provides a way to automatically determine the number of clusters based on standard statistics, taking outlier or noise into account. It therefore yields robust clustering methods.

Constraint-based Method

In this method, the clustering is performed by the incorporation of user or application-oriented constraints. A constraint refers to the user expectation or the properties of desired clustering results. Constraints provide us with an interactive way of communication with the clustering process. Constraints can be specified by the user or the application requirement.

Classification

Classification is a data mining technique that assigns categories to a collection of data in order to aide in more accurate predictions and analysis. Also called sometimes called a Decision Tree, classification is one of several methods intended to make the analysis of very large data sets

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
http://www.ijcsjournal.com Volume 5, Issue 1, No 25, 2017 ISSN: 2348-6600
Reference ID: IJCS-280 PAGE NO: 1819-1829
Alagappa University, Karaikudi, India 15th -16th February 2017
IT Skills Show & International Conference on Advancements in Computing Resources (SSICACR-2017)
http://aisdau.in/ssicacr ssicacr2017@gmail.com

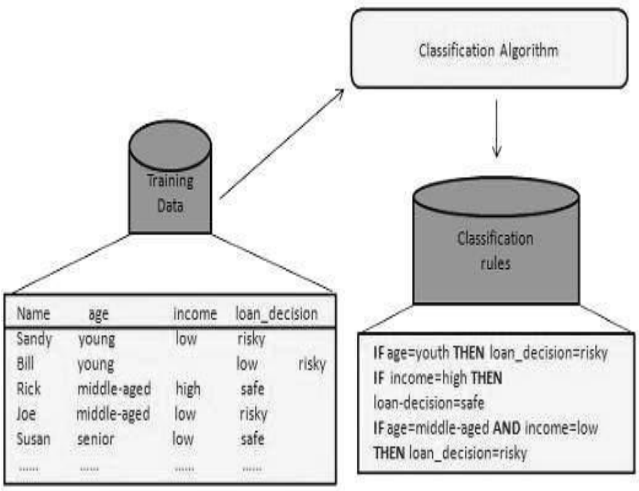
effective.Classification models predict categorical class labels

The Data Classification process includes two steps –

- Building the Classifier or Model
- Using Classifier for Classification

Building the Classifier or Model

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

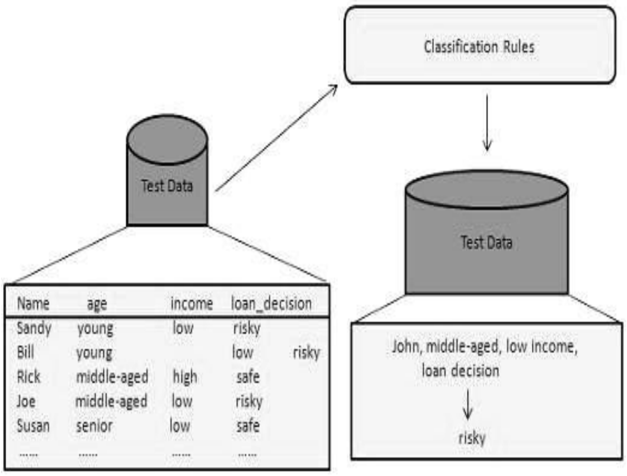


Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

IF age=youth THEN loan_decision=risky
IF income=high THEN loan-decision=safe
IF age=middle-aged AND income=low THEN loan_decision=risky

Using Classifier for Classification

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.



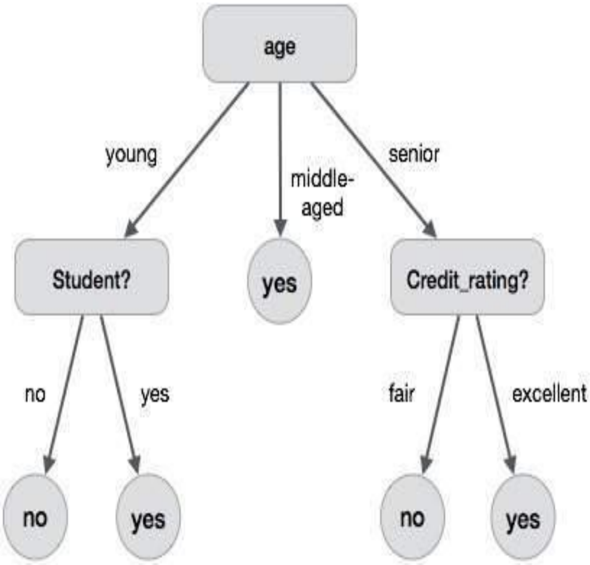
Name	age	income	loan_decision	
Sandy	young	low	risky	
Bill	young		low	risky
Rick	middle-aged	high	safe	
Joe	middle-aged	low	risky	
Susan	senior	low	safe	
......	......	......	......	

John, middle-aged, low income, loan decision
↓
risky

Methods of classification:

A decision tree is a structure that includes a root node, branches, and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test, and each leaf node holds a class label. The topmost node in the tree is the root node.

The following decision tree is for the concept buy_computer that indicates whether a customer at a company is likely to buy a computer or not. Each internal node represents a test on an attribute. Each leaf node represents a class.



The benefits of having a decision tree are as follows −

- It does not require any domain knowledge.
- It is easy to comprehend.
- The learning and classification steps of a decision tree are simple and fast.

Decision Tree Induction Algorithm

A machine researcher named J. Ross Quinlan in 1980 developed a decision tree algorithm known as ID3 (Iterative Dichotomiser). Later, he presented C4.5, which was the successor of ID3. ID3 and C4.5 adopt a greedy approach. In this algorithm, there is no backtracking; the trees are constructed in a top-down recursive divide-and-conquer manner.

Tree Pruning

Tree pruning is performed in order to remove anomalies in the training data due to noise or outliers. The pruned trees are smaller and less complex.

Tree Pruning Approaches

Here is the Tree Pruning Approaches listed below –

- **Pre-pruning** − The tree is pruned by halting its construction early.
- **Post-pruning** - This approach removes a sub-tree from a fully grown tree.

Cost Complexity

The cost complexity is measured by the following two parameters −

- Number of leaves in the tree, and

- Error rate of the tree.

Bayesian classification method:

Bayesian classification is based on Bayes' Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.

Baye's Theorem

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities −

- Posterior Probability [P(H/X)]
- Prior Probability [P(H)]

where X is data tuple and H is some hypothesis.

According to Bayes' Theorem,

P(H/X)= P(X/H)P(H) / P(X)

Bayesian Belief Network

Bayesian Belief Networks specify joint conditional probability distributions. They are also known as Belief Networks, Bayesian Networks, or Probabilistic Networks.

- A Belief Network allows class conditional independencies to be defined between subsets of variables.
- It provides a graphical model of causal relationship on which learning can be performed.
- We can use a trained Bayesian Network for classification.

There are two components that define a Bayesian Belief Network −

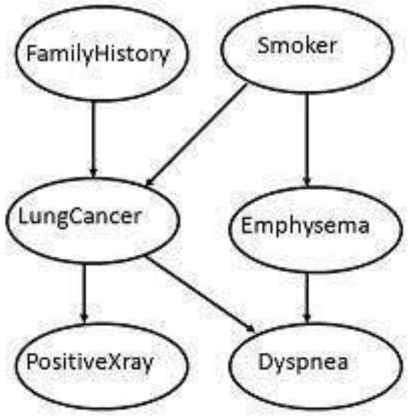
- Directed acyclic graph
- A set of conditional probability tables

Directed Acyclic Graph

- Each node in a directed acyclic graph represents a random variable.
- These variable may be discrete or continuous valued.
- These variables may correspond to the actual attribute given in the data.

Directed Acyclic Graph Representation

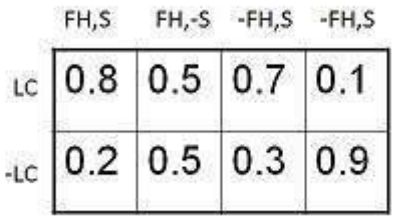
The following diagram shows a directed acyclic graph for six Boolean variables.



The arc in the diagram allows representation of causal knowledge. For example, lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker. It is worth noting that the variable PositiveXray is independent of whether the patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has lung cancer.

Conditional Probability Table

The arc in the diagram allows representation of causal knowledge. For example, lung cancer is influenced by a person's family history of lung cancer, as well as whether or not the person is a smoker. It is worth noting that the variable PositiveX-ray is independent of whether the patient has a family history of lung cancer or that the patient is a smoker, given that we know the patient has lung cancer.



	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

IF-THEN Rules

Rule-based classifier makes use of a set of IF-THEN rules for classification. We can express a rule in the following from –

IF condition THEN conclusion

Let us consider a rule R1,

R1: IF age=youth AND student=yes
THEN buy_computer=yes

NODE:

- The IF part of the rule is called rule antecedent or precondition.
- The THEN part of the rule is called rule consequent.
- The antecedent part the condition consist of one or more attribute tests and these tests are logically ANDed.
- The consequent part consists of class prediction.

We can also write rule R1 as follows:

R1: (age = youth) ^ (student = yes))(buys computer = yes)

If the condition holds true for a given tuple, then the antecedent is satisfied.

Rule Extraction

Here we will learn how to build a rule-based classifier by extracting IF-THEN rules from a decision tree.

NODE

- One rule is created for each path from the root to the leaf node.
- To form a rule antecedent, each splitting criterion is logically ANDed.
- The leaf node holds the class prediction, forming the rule consequent.

Rule Pruning

The rule is pruned is due to the following reason –

- The Assessment of quality is made on the original set of training data. The rule may perform well on training data but less well on subsequent data. That's why the rule pruning is required.
- The rule is pruned by removing conjunct. The rule R is pruned, if pruned version of R has

greater quality than what was assessed on an independent set of tuples.

FOIL is one of the simple and effective method for rule pruning. For a given rule R,

FOIL_Prune = pos - neg / pos + neg

wherepos and neg is the number of positive tuples covered by R, respectively.

Note − This value will increase with the accuracy of R on the pruning set. Hence, if the FOIL_Prune value is higher for the pruned version of R, then we prune R.

COMPARATIVE FOR CLUSTER AND CLASSIFICATION:

S.no	CHARACTER	CLUSTER	CLASSIFICATION
1	Definition	Group of similar (or) dissimilar data object.	Assign the predefined tags for the object.
2	Supervision	Unsupervised learning technique	Supervised learning technique
3	Process	Datasets are split into subsets with similar features.	Categorize the new data according to the observations of the training set.
4	Training set	A training set is not used	A training set is used to find similarities
5	Lables	No labels	Labels for some points.
6	Aim	To find whether there is any relationship between them.	To find which class a new object belongs to from the set of predefined classes.

7	Application	Data Mining Pattern recognition Image analysis Bioinformatics Machine Learning Voice mining Image processing Text mining Web cluster engines Whether report analysis	Email spam Bank Cancer tumour cell identification Drug classification Facial key points Automotive car driving
8	Algorithm	Partitioning Method, Hierarchical Method, Density-based Method, Grid-Based Method, Model-Based Method, Constraint-based Method	Decision tree, rule based method, Bayesian algorithm
9	Advantage	Automatic recovery from failure	Easy to interpet,easy to implement,fast application
10	Disadvantage	Clustering are complexity and inability to recover from database corruption	Simple decision boundaries,problems with lots of missing data

CONCLUSION

Data mining is covering every field of our life. Mainly we are using the data mining in banking, education, business etc. In this paper, we have provided an overview of the comparison of clustering algorithms such as partitioning, hierarchical density based and grid based methods and classification methods of decision tree, rule based ,Bayesian algorithm.K-mean algorithm has biggest advantage of clustering. Different classification algorithm gives different result on baseof accuracy,training time,precision . The clustering techniques provides consistent results in the domain of bioinformatics, text mining, image processing and web cluster engines whereas classification greatly help us in e-commerce and e-transaction domains to categories the patterns of transaction. Thus each of this technique have their unique characteristics to produce quality results.

References

[1]Jiawei Han and M Kamber , Data Mining: Concepts and Techniques, Second Edition.

[2]B.G.Obula Reddy, Dr. MaligelaUssenaiah, ìLiterature Survey on Clustering Techniquesî, IOSR Journal of Computer Engineering

[3]S.Neelamegam, Dr.E.Ramaraj, “Classification algorithm in Data mining: An Overview”, International Journal of P2P Network Trends and Technology, Volume 4 Issue 8-Sep 2013

[4]S.Archana1, Dr. K.Elangovan, “Survey of Classification Techniques in Data Mining”,International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February-2014

[5]Dr.A.Bharathi, E.Deepankumar ,” Survey on Classification Techniques in Data Mining”,International
Journal on Recent and Innovation Trends in Computing and Communication, Volume. 2 Issue. 7,July 2014