A DETAILED STUDY ON CLASSIFICATION TECHNIQUES FOR DATA MINING

[1]Jyoti Kesarwani, [2]Kshama Tiwari

[1]M. Tech Student, UIT College, Allahabad Uttar Pradesh
[2]Assistant Professor, UIT College, Allahabad, Uttar Pradesh

---

Abstract – *Extraction of useful information from huge amount of data is known as data mining also known as knowledge discovery in database (KDD). There are so many sources that generates data in a very large amount like social networking sites, camera, sensors etc. This is the main reason that data mining is increasing rapidly. This paper presents a survey of clustering techniques and tools used for data mining. Classification is a supervised learning technique in which it identifies the class of unknown objects whereas clustering is an unsupervised learning. Clustering is the process of partitioning a set of data objects into subsets. Objects with in a cluster are more similar and dissimilar to other clusters. The similarity between objects is calculated using various distance measures like Euclidean distance, Manhattan distance, cosine etc.*

***Key Words*: Data Mining, Machine Learning, Classification, clustering algorithms, Supervised, Unsupervised Learning**

1. INTRODUCTION

Data mining plays a very important role for finding the frequent data pattern from internet, data set, data warehouse, data mart etc. Data mining, also called as **data archeology, data dredging, data harvesting**, is the process of extracting hidden knowledge from large volumes of raw data and using it to make critical business decisions. Data mining is used in various applications like finance, marketing, banking, credit card fraud detection, whether prediction.
Data mining helps to extract hidden patterns and make hypothesis from the raw data. Data mining process has mainly 7 steps as Data integration, data cleaning, data selection, data transformation, data mining, pattern evaluation and knowledge representation [1]. This process is shown in Fig-1.

Data Cleaning: Data in the real world is dirty, means incomplete, noisy and inconsistent data. Quality decisions must be based on quality data. So, before performing the analysis on the raw data, data cleaning is performed, which includes the following tasks:

- Filling missing values.
- Smooth noisy data and remove outliers by using algorithms like Binning algorithm.
- Resolve inconsistencies.

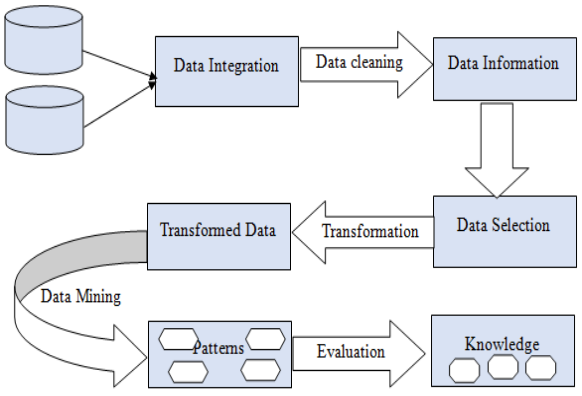

Figure 1: Data Mining Process

Data Integration: where multiple heterogeneous data sources may be combined.

Data Selection: Where task relevant data are selected from data warehouse or any other data sources including www, RDBMS etc.

Data Transformation: In data transformation, the data are transformed into format appropriate for data mining. For ex: An attribute data may be normalized so as to fall between a small range 0 to 1. It includes the following tasks:

- **Smoothing:** which works to remove noise from the data. Such techniques include binning, regression and clustering.
- **Aggregation:** Various aggregation operations such as mean and median are applied to the data. For ex: the daily sales data may be aggregated.
- **Normalization:** where the attribute data are scaled so as to fall within a small specified range, such as 0 to 1.

Data Mining: It is the process of extraction of interesting information or patterns from data in large database is known as data mining.

Pattern Evaluation: This component typically employs interestingness measures and interacts with the data mining modules so as to focus the search towards interesting patterns.

Knowledge representation: Various visualization and knowledge representation techniques are used to present the extracted knowledge to the user.

2. Related Work

A lot of researchers have implemented various data mining approaches in order to solve the various problems related to forecasting and analysis. Decision tree algorithm is a kind of data mining model to make induction learning algorithm based on examples. It is easy to extract display rule, has smaller computation amount, and could display important decision property and own higher classification precision. We select decision-making tree which is very visible and easy realized as data mining tools, and set up decision-making tree model which is used to predict groups of elements. Nowadays is necessary to take decisions based in the knowledge obtained through advanced techniques of date analysis, decision tree is an interesting option. In this work a Rich Internet Application to visualize a decision tree in a mobile device is presented. This application lets deploy the complete tree decision and the categorization of new registers, with this tool is possible to take decisions based in the analysis of data in an extended data base. The support vectors play an important role in the training to find the optimal hyper-plane. For the problem of many non-support vectors and a few support vectors in the classification of SVM, a method to reduce the samples that may be not support vectors is proposed in this paper. First, adopt the Support Vector Domain Description to find the smallest sphere containing the most data points, and then remove the objects outside the sphere. Second, remove the edge points based on the distance of each pattern to the centers of other classes *k*-nearest neighbor algorithm (*k*NN) which usually identifies the same number of nearest neighbors for each test example. It is known that the value of *k* has crucial influence on the performance of the *k*NN algorithm, and our improved *k*NN algorithm focuses on finding out the suitable *k* for each test example. The proposed algorithm finds out the optimal *k*, the number of the fewest nearest neighbors that every training example can use to get its correct class label. For classifying each test example using the *k*NN algorithm, we set *k* to be the same as the optimal *k* of its nearest neighbor in the training set. Naive bayes classifier, a classification method based on bayes theory, shows excellent properties in many fields.

3. Classification Algorithms

3.1 Decision Tree Induction:

Decision tree induction is the learning of decision trees from class labeled tuples. A decision tree is a flow chart like tree structure where each internal node denotes a test on an attributes, each branch represents an output of the test and each leaf node denotes a class label. Decision trees are trees that classify data by sorting them based on feature values. These decision tree induction methods are supervised machine learning methods that construct decision tree from a set of input output values. A decision tree uses top down approach that searches solution from search spaces.

In order to classify an unknown sample, the attribute values of the sample are tested against the decision tree. A path is traced from the root to a leaf node that holds the class prediction for that sample. Decision trees were then converted to classification rules using IF-THEN-ELSE.
A typical Decision Tree is shown in Figure 1. This represents the concept buy a computer that is, the tree tries to predict whether a customer of an electronics shop or cannot buy a computer. The internal nodes are denoted by rectangles and leaf nodes ovals are denoted by [3].

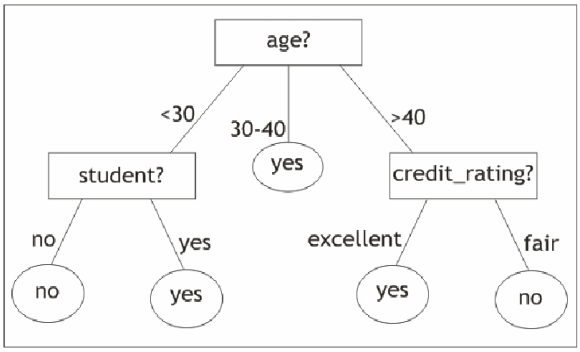

Figure 2: Decision Tree Example

3.2 K Nearest Neighbour Algorithm:

KNN means k nearest neighbor. It is a very simple algorithm. Given N training vectors, suppose we have 'a' and 'o' letters as training vectors in the bi dimensional feature space. the KNN algorithm identifies k nearest neighbors of 'c'.'c' is another space vector that we want to estimate its class regardless of labels.
The kNN expects the class conditional probabilities to be locally constant, and suffers from bias in high dimensions. *k*NN is an extremely flexible classification scheme, and does not involve any preprocessing of the training data. This can offer both space and speed advantages in very large problems.

KNN is an example-based learning group. This algorithm is also one of the lazy learning techniques. KNN is done by searching for the group of K objects in the closest training data (similar) to objects in new data or data testing [2]. Generally, the Euclidean distance formula is used to define the distance between two training objects and testing [10].

$$dxy = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

3.3 Naïve Bayes Classification:

"naive" Bayes classification is a method of supervised learning if the attributes are conditionally independent given the classes.

It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as 'Naive'.

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods.

It tells us how often A happens *given that B happens*, written **P(A|B)**, when we know how often B happens *given that A happens*, written **P(B|A)** , and how likely A and B are on their own.

- P(A|B) is "Probability of A given B", the probability of A given that B happens

- P(A) is Probability of A
- P(B|A) is "Probability of B given A", the probability of B given that A happens
- P(B) is Probability of B

3.4 Support Vector Machine(SVM)

"Support Vector Machine" (SVM) is a supervised machine learning algorithm which can be used for classification. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes

Binary SVM:

Binary classification is a technique to find the category of data points.
For example- Let us consider that C1 and C2 are the two class labels. And we have data point one is positive and other is negative but here the problem is they are mixed so we need no find the decision boundary between the class label and support vectors. There could be exist more than one separable line but we need to identify the maximum margin line from the support vectors and this line is called 'Decision Boundary'. And one side of decision boundary are positive points and other side has negative points.

4. Comparison of Different Classification algorithms

This section discusses the comparison between various classification algorithms with their advantages and disadvantages. Table I provides information about various algorithms.

Table I: Comparison of Classification algorithms

Algorithm	Findings	Advantages	Disadvantages
Decision Tree	Decision tree is a supervised learning method to construct trees from a set of input output samples.	It is simple to understand, interpret and have little effort form user for data preparation. Easy to determine worst, best and expected values for different scenarios.	If we do small change in the data can lead to a large change in the structure of the optimal decision tree. Calculations can get complex, if values are uncertain and/or if many outcomes are linked.
SVM	SVM is a supervised learning in which we plot each data item as a point in n-dimensional space. with the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes .	It works really well with clear margin of separation It is effective in high dimensional spaces.	It doesn't perform very well, when target classes are overlapping SVM doesn't directly provide probability estimation.
Naïve Bayes	The Naïve Bayes Classification represents a supervised learning method as well as a statistical method for classification. It is probabilistic model and it allows us to find uncertainty and determining probabilities of the outcomes. It can solve diagnostic and predictive .	It is very simple, easy to implement and fast. It can make probabilistic predictions. It handles both continuous and discrete data.	Naive Bayes classifier make assumption on the shape of your data distribution, i.e. any two features are independent given the output class.

5. Conclusion:

In this paper, we have presented the survey of various classification algorithms used for analysis. There are mainly three types of classification methods are discussed.

REFERENCES:

[1] Kiran Kumar Patro, P. Rajesh Kumar, "Denoising of ECG Raw signal by cascaded window based digital filters configuration", IEEE Power, Communication and Information Technology Conference (PCITC), Oct, 2015.

[2] Bhumika Chandrakar, O.P.Yadav and V.K.Chandra, "A survey of noise removal techniques for ECG signal", Int. Journal of Advanced Research in Computer and Communication Engineering, March 2013.

[3] Mostafa Guda, Safa Gasser, "MATLAB Simulation Comparison for Different Adaptive Noise Cancelation Algorithms", the SDIWC in 2014.

[4] Sarita Mishra, Debasmit Das, Roshan Kumar and Parasuraman Sumathi, " A power-line intereference canceler based on sliding DFT Phase locking scheme for ECG signals", IEEE Transactions on Instrumentation & Measurement, Vol.64, No.1, Jan 2015.

[5] Prakruti J.joshi, Vivek P.Patkar, "ECG denoising using MATLAB" Int. Journal of Scientific & Engineering Research, May-2013.

[6] Mbachu C.B. , Offor K.J, "Reduction of power line noise in ECG signal using FIR digital filter implemented with hamming window",Int. Journal of Science, Environment and Technology, 2013.

[7] Fatin A. Elhaj, Naomie Salim, Arief R. Harris, Tan Tian Swee,Taqwa Ahmed, "Arrhythmia recognition and classification using combined linear and nonlinear features of ECG signals", Computer Methods and Programs in Biomedicine I27, Elsevier Ireland Ltd., Pg 52-63, 2016.

[8] Aya F. Khalaf , Mohamed I. Owis, Inas A. Yassine, "A novel technique for cardiac arrhythmia classification using spectral correlation and support vector machines", Expert Systems with Applications 42, Elsevier Ltd., Pg 8361-8368, 2015.

[9] Sakuntala Mahapatra, Debasis Mohanta, Prasant Mohanty, Santanu kumar Nayak, Pranab kumar Behari, "A Neuro-fuzzy based model for analysis of an ECG signal using Wavelet Packet Tree", 2nd International Conference on Intelligent Computing, Communication & Convergence, ICCC-2016, Elsevier Ltd, Odisha, India, Pg 175-180.

[10] Dae-Geun Jang, Seung-Hun Park, and Minsoo Hahn, "A Gaussian Model-Based Probabilistic Approach for Pulse Transit Time Estimation", IEEE Journal of Biomedical and Health Informatics,Vol.20, No.1, Jan 2016.

[11] Raquel Gutiérrez-Rivas, J. Jesús García, William P. Marnane, and Alvaro Hernández, "Novel Real-Time Low-Complexity QRS Complex Detector Based on Adaptive Thresholding", IEEE Sensors Journal, VOL. 15, NO. 10, October 2015.

[12] Michael Alb, Piergiorgio Alotto1, Christian Magele, Werner Renhart, Kurt Preis and Bernhard Trapp, "Firefly Algorithm for Finding Optimal Shapes of Electromagnetic Devices", IEEE Transactions On Magnetics, VOL. 52, NO. 3, March 2016.

[13] Jyh-Shing and Roger Jang., "ANFIS: AdaptiveNetwork-Based Fuzzy Inference System," computer methods and programs in biomedicine, IEEE Transactions on Systems, University of California,1993

[14] Abdulkadir Sengur., "An expert system based on linear discriminant analysis and adaptive neurofuzzy inference system to diagnosis heart valve diseases," Expert Systems with Applications, 2008.

[15] G. Zhao, C. Peng and Xiting Wang., "Intelligent Control for AMT Based on Driver's Intention and ANFIS Decision-Making," World Congress on Intelligent Control and Automation, 2008.

[16] Anupam Das, J. Maiti and R.N. Banerjee., "Process control strategies for a steel making furnace using ANN with bayesian regularization and ANFIS," Expert Systems with Applications, 2009.

[17] N. Deepak, Anu Mathew, "Adaptive Neuro-Fuzzy Inference System for Classification of ECG Signal", International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering,
Vol. 1, Issue 1, July 2012.