Survey on Classification Techniques in Data Mining

Dr. A. Bharathi¹, E. Deepankumar²
¹Professor, Department of IT, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India. bharathia@bitsathy.ac.in.
²PG Scholar Department of IT, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India. deepankumar.se13@bitsathy.ac.in.

Abstract—Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, different, useful and understandable patterns in large databases. Classification is a data mining technique used to predict group membership for data instances. In this paper, we present the data mining tasks and classification techniques. Different kinds of classification techniques discussed. Those techniques are Association Rule Mining, Bayesian Classification, and Decision Tree Classification, nearest neighbor classifier, neural Networks and Support Vector Machine.

Keywords: Data Mining Tasks, Classification, rule mining, classification, support vector machine

I. Introduction

Data mining is a process of extracting or mining the useful pattern or information and relationships in the large amounts or massive volumes of data. The term data mining is also called as ‘Knowledge mining from data’ or “Knowledge mining”. [1][3] the goal of the Data mining to discover previously unknown association among data, usually when data come from various data bases.

Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. [2] The Knowledge Discovery from Data or KDD process steps is given below [1] in fig 1.

Fig. 1: KDD Process

1. Data Cleaning
   To eliminate noise and inconsistent data.
2. Data Integration
   Multiple data sources may be combined.
3. Data Selection
   The data consistent to the test work are retrieved from the database.

4. Data Transformation
   Data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations.
5. Data Mining
   Examining the large pre existing data bases in order to generate new information.
6. Pattern Evaluation:
   To identify the truly interesting ting knowledge based on interestingness measures.
7. Knowledge presentation
   The visualization and knowledge representation techniques are used to present mined knowledge to users.

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. Data mining tasks can be classified in two categories-descriptive and predictive. [1][2][3] Descriptive mining or clustering, also divides data into groups. With clustering, however, the proper groups are not known in advance; the patterns discovered by analyzing the data are used to determine the groups. [15] Predictive mining is used when the goal is to estimate the value of a particular target attribute and there exist sample training data for which values of that attribute are known.

This paper is organized as follows: Section II describes Data mining Tasks, Section III Classification Techniques, Section IV Conclusion, and Section V References.

II. Data Mining Tasks

Data mining used in different type of techniques to extract the knowledge from the data, the techniques are: [3]

- Anomaly detection (Outlier/ deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.(ex-bank fraud identification)
- Three type of anomaly detections: [15]
Unsupervised Anomaly Detection techniques detect anomalies in an unlabeled test data set under the assumption
Supervised Anomaly Detection
Techniques require a data set that has been labeled as "normal" and "abnormal" and involves training a classifier
Semi-Supervised Anomaly detection techniques construct a model representing normal behavior from a given normal training data set, and then testing the likelihood of a test instance to be generated by the learnt model

- Association rule learning (Dependency modeling) – Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

In this data mining techniques used to mining the different kind of data, the forms of data for mining applications are database data, data warehouse data, transactional data, data streams, graph or networked data, spatial data, text data, multimedia data, and the World Wide Web data.

III. Classification Techniques

Classification is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. [5][6]

- Association rule mining
- Bayesian Classification
- Decision tree classification
- Nearest Neighbor
- Neural Networks(Back Propagation)
- Support Vector Machines (SVMs)

3.1 Association Rule Mining

Association rule mining, one of the most extensive and well researched techniques of data mining,[3] It aims to extract unusual relationship, frequent patterns, associations or causal structures among sets of items in the transaction databases or other data repositories. [9] Association rules are widely used in various areas such as telecommunication networks, market and risk management, inventory control etc. Various association mining techniques and algorithms will be briefly introduced and compared later. [10] Association rule mining is to find out association rules that satisfy the predefined minimum support and confidence from a given database. The problem is usually decomposed into two sub problems. One is to find those item sets whose occurrences exceed a predefined threshold in the database; those item sets are called frequent or large item sets. The second problem is to generate association rules from those large item sets with the constraints of minimal confidence.

3.2 Bayesian Classification

A naive Bayes classifier assumes that the presence or absence of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. [11] A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of the presence or absence of the other features. [8] Naive Bayes classifiers can be trained very efficiently in a supervised learning setting. An advantage of naive Bayes is that it only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. [14] Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

3.3 Decision Tree Classification

Decision tree classification approach is most useful in classification problems. [3] It is a flow chart like tree structure. Trees are constructed in a top down recursive divide and conquer manner. [15] In this classification method used in different type algorithm to classify the data sets, the algorithms are: [1]

- ID3(Iterative Dichotomiser)
- C4.5(a Successor of ID3)
- Classification and Regression Trees(CART)

The algorithm follows a top-down approach, which starts with a training set of tuples and their associated class labels.

Advantages

Rules can be generated that are easy to interpret and understand. It is scalable for large database because the tree size is independent of the database size. [3] Each tuple in the database must be filtered through the tree, and time is proportional to the height of the tree.

Disadvantages

It is does not handle continuous data. [3] Handling missing data is difficult because correct branches in tree could not be taken the labels.

3.4 Nearest Neighbor

Nearest neighbor classifiers are based on learning by analogy. The training samples are described by n dimensional numeric attributes [4]. Each sample represents a point in an n-dimensional space. In this way, all of the training samples are stored in an n-dimensional pattern
space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. [9] The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. Nearest neighbor classifiers are instance-based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified.[15]

3.5 Neural Networks

Neural network represent a brain image or symbol for Information processing.[1][3] These models are biologically inspired rather than an exact replica of how the brain actually functions.[8] Neural networks have been shown to be very talented systems in many forecasting applications and business classification applications due to their ability to “learn” from the data, their nonparametric nature (i.e., no rigid assumptions), and their ability to generalize. Neural computing refers to a pattern recognition methodology for machine learning.[11] The resulting model from neural computing is often called an artificial neural network (ANN) or a neural network. Neural networks have been used in many business applications for pattern recognition, forecasting, prediction, and classification.

The human brain possesses distract capabilities for information processing and problem solving that modern computer cannot compete with in many aspects. [1]It has been predicate that a model or a system that is open minded or liberal and supported by the results from brain research, with a structure similar to that of biological neural networks, could exhibit similar intelligent functionality. Based on this bottom-up guess, ANN (also known as connectionist models, parallel distributed processing models, neuromorphic systems, or simply neural networks) have been developed as biologically inspired and plausible models for various tasks. [12] Biological neural networks are composed of many massively interconnected primitive biological neurons. Each neuron possesses axons and handwrites finger-like projections that enable the neuron to communicate with its neighboring neurons by transmitting and receiving electrical and chemical signals. More or less resembling the structure of their counterparts, ANN are composed of interconnected, simple processing elements called artificial neurons. ANN possess some desirable traits similar to those of biological neural networks, such as the capabilities of learning, self-organization, and fault tolerance.

3.6 Support Vector Machine

SVM has develop as one of the most popular and useful techniques for data classification [7]. It can be used for classify the both linear and non linear data. [2] The objective of SVM is to produce a model that predicts the target value of data occurrence in the testing set in which only attributes are given.[8] The classification goal in SVM is to separate the two classes by means of a function prepare from available data and thereby to produce a classifier that will work well on further unseen data. [8] The simplest form of SVM classification is the maximal margin classifier. It is used to solve the most basic classification problem, namely the case of a binary classification with linear separable training data. [1] The aim of the maximal margin classifier is to find the hyperplane with the largest margin, i.e., the maximal hyperplane, in real-world problems, training data are not always linear separable. In order to handle the nonlinerly separable cases some slack variables have been introduced to SVM so as to tolerate some training errors, with the influence of the noise in training data thereby decreased. This classifier with slack variables is referred to as a soft-margin classifier.

IV Conclusion

Data mining offers promising ways to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. In this paper, we present the basic classification techniques. Several major kinds of classification method including Association Rule Mining, Decision tree classification, Bayesian Classification, nearest neighbour classification, Neural Network, and Support Vector Machine.

V. References

[1] Jiawei Han, Micheline Kambar, Jian Pei, “Data Mining Concepts and Techniques” Elsevier Second Edition.

