Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.241 – 248

RESEARCH ARTICLE

A Study on Different Classification Models for Knowledge Discovery

Neha Gulia[1], Sugandha Singh[2], Luxmi Sapra[3]
[1]Computer Science, PDMCE Bhadurgarh, India
[2]Computer Science, PDMCE Bhadurgarh, India
[3]Computer Science, PDMCE Bhadurgarh, India
[1] gulia1neha@gmail.com, [3] luxmi_engg.pdm.ac.in

Abstract: Knowledge discovery is the major data processing activity applied to derive the hidden and valuable information. Classification is one of such information mining activity applied in various applications areas. In this paper, a study on classification model is defined along with associated approaches. The paper has described in this classification model in generalized form as well as relative to other knowledge discovery methods.

Keywords: Clustering, Classification, KDD, Supervised

I. INTRODUCTION

Data mining process includes understanding the business requirements and needs. While understanding the business requirement both data and business requirements are understood. Then, using this business requirement it identifies data source and data format in this the data is prepared and modeled for evaluation, and then using these data source and data format it build data model. This data model is used to build data structure. Then, the mining operation is performed on this data structure.

Data mining field comprises of four main disciplines:

Statistics: defines tools for measuring significance in the data
Machine learning: provide algorithm to induce knowledge from the data
Artificial intelligence: involve knowledge for encoding and search techniques.

Data management and databases: provides an efficient way of accessing and maintaining data. Data mining has many well known tasks and one which is very interesting to study and analyze is categorization.

The categorization is a supervised form of machine learning. Machine learning comprises of supervised, unsupervised, semi- supervised and reinforced learning. In the supervised form of learning the learning is from the training data available.

Machine learning is a branch of artificial intelligence. It's the construction and study of the systems that can learn from the system. Arthur Samuel (1959) defined machine learning a field of study that gives computer ability to learn without being explicitly programmed.

There are two machine learning algorithms:

- Supervised Learning
- Unsupervised Learning

A) Knowledge Discovery in Databases

Data Mining is the core of KDD process. KDD is concerned with the development of method and techniques for making sense of data. KDD process is the traditional method of turning data into knowledge on manual analysis.
KDD Process is process of finding knowledge in data & is the high level application of data mining [7]. It maps low-level data into other form. Knowledge discovery in databases is the non-trivial process. The knowledge includes the intelligent system process that defines the basic data modeling process stages.

The basic difference between KDD process & Data Mining is that DM is one of the steps of the KDD process. DM analysis the data and KDD process discover the useful knowledge from data given. Knowledge discovery consists of an iterative sequence of following steps:

1) Understand your goal or domain.
2) Then create the dataset and select it.
3) Clean the selected dataset and transformed into appropriate form for mining.
4) Then apply the intelligent methods (DM Methods) on transformed dataset in order to extract data patterns.
5) When patterns are obtained evaluation, interpret and visualization is done to identify the patterns representing knowledge are
6) At the end Knowledge presentation is done to present the knowledge to the user and manage the discovered knowledge.

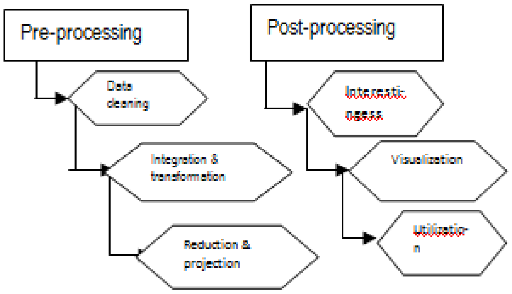


Fig 1.1 KDD Processes

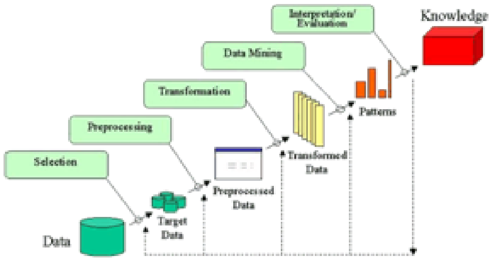


Fig 1.2 Architecture of KDD process

The Knowledge Discovery in Databases (KDD) field of data mining is concerned with the development of methods, techniques and algorithm which can make sense of the available data. KDD is useful in finding trends, patterns, correlations and anomalies in the databases which is helpful to make accurate decisions for the future. Association rule mining finds collections of data attributes that are statistically related to the data.

B) Association Rule Mining

Association rule mining is an important technique in the data mining. A major concern today in Association rule mining is to improve the algorithmic performance. Association rule mining is to discover the potential relation between the sets of data items. Many algorithms were defined under association mining they are Eclat algorithm, FP-growth algorithm, GUHA procedure ASSOC, OPUS search.

In the algorithms of association mining, Apriori is the oldest which is offered by Agrawal R in 1993. Apriori is the best one under association mining. It uses a breadth first search technique. Eclat algorithm uses set interaction and is a depth first search technique. In the FP - growth algorithm, FP stands for Frequent pattern, and uses recursive processing approach. GUHA procedure ASSOC uses fast bit operation and is a method for exploratory data analysis. OPUS is an efficient association technique but does not require monotone, anti-monotone constraint.

Other categories of association rule mining were contrast set learning, weighted class learning, high-order pattern discovery, K-optimal pattern discovery, generalized association rules, quantitative association rules, interval Data association rules, maximal association rules, sequential pattern mining, sequential Rules.

Association rule mining was used to find association and interesting association between the data in the large data set.

Basically, these association rule mining algorithm were defined as class association rules (CAR) mining. This CAR comprises of three steps:

- Rule generation: rules are generated from the dataset.
- Rule ordering: classifiers are used to order the rules. Ordering can be done according to many criteria's. Some of which are used in a priori are supported, confidence, minimum support threshold, rule length, etc.
- Classification: In this step, according to the rule specified the data are classified.

C) Classification

Classification is an important area of research in data mining. Classification partitions massive quantities of data into sets of common characteristics and properties[9].classification a set of records, acting as *training set,* is analysed in order to produce a model of the given data. Each record is assumed to belong to a predefined class, as determined by one of the attributes, called *attribute.* Table 1 shows a part from a sample training set of the medical database, where each record represents a patient and *Lived* is the *classifying attribute* of the training set.

- Once derived, the classification model can be used to categories future data samples, as well as
- providing a better understanding of the database contents. Classification is particularly useful when a
- database contains examples that can be used **as** the basis for future decision making, e.g. for
- assessing credit risks, for medical diagnosis, or for scientific data analysis.

The classification technique that developed in Easy Miner is based on the decision tree structure. By using a decision tree, untagged data sample can be classified by testing the attribute values of the sample data against the decision tree. **A** path is produced from the root to a leaf which has the class identification of the sample.

II. LITERATURE SURVEY

In Year 2002, Adepele Olukunle performed a work," A Fast Algorithm for Mining Association Rules in Medical Image Data". This paper presents a fast association rule mining algorithm which is suitable for medical image data sets. Author provide a flavour of Presented implementation environment. Author also give an example, how Presented proposed algorithm work to assess its suitability.

In Year 2006, Carlos Ordonez performed a work," Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction". Association rules represent a promising technique to improve heart disease prediction. Author introduce an algorithm that uses search constraints to reduce the number of rules, searches for association rules on a training set, and finally validates them on an independent test set. The medical significance of discovered rules is evaluated with support, confidence, and lift. Association rules are applied on a real data set containing medical records of patients with heart disease. In medical terms, association rules relate heart perfusion measurements and risk factors to the degree of disease in four specific arteries.

In Year 2008, Chunxue Shi performed a work," Path Planning for Deep Sea Mining Robot Based on ACO-PSO Hybrid Algorithm". In this study, the environment model was established by Bitmap method, and robot movement was simplified into particle movement by using Framework Space method. Ant colony optimization (ACO) is used to establish the corresponding solution, and some material algorithm steps are set out. Particle swarm optimization (PSO) is applied to optimize the parameters

In Year 2009, Gaurav N. Pradhan performed a work," ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA". In this paper, Author consider real-life time series data of muscular activities of human participants obtained from multiple Electromyogram (EMG) sensors and discover patterns in these EMG data streams. Each EMG data stream is associated with quantitative attributes such as energy of the signal and onset time which are required to be mined along with EMG time series patterns. Author propose a two-stage approach for this purpose: in the first stage, Presented emphasis is on discovering frequent patterns in multiple time series by doing sequential mining across time slices. Presented evaluation with large sets of time series data from multiple EMG sensors demonstrate that Presented two-stage approach speeds up the process of finding association rules in such multidimensional environment as compared to other methods and scales up linearly in terms of number of time series involved. Presented approach is generic and applicable to any multiple time series dataset format.

Mr.K.Ravikumar performed a work," ACO based spatial Data Mining for Traffic Risk Analysis". There was a first study aiming identifying and at predicting the accident risk of the roads. It used a decision tree that learns from the inventoried accident data and the description of the corresponding road sections. The existing work provided a pragmatic approach to multilayer geo-data mining. The process behind was to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build' a decision tree. Presented method has higher efficiency in performance of the discovery process and in the quality of trend patterns discovered compared to other existing approaches using non-intelligent decision tree heuristics.

In Year 2010, Wei Wang performed a work," Mining Association rules in Medical Data Based on Concept Lattice". paper introduces some definitions of concept lattice, and compares two methods of data inductive: AOI and concept lattice. After introduce the context, actual medical data is discretized and concept lattice and Hasse diagram are constructed to generate the concept hierarchy, and finally some strategies is used to extract interesting association rules.

In Year 2010, Mostafa Fathi Ganji performed a work," Parallel Fuzzy Rule Learning Using an ACO-Based Algorithm for Medical Data Mining". This paper proposes a rule-based system for medical data mining by using a combination of ACO and fuzzy set theory, named FACO-Miner. Author have proposed a new heuristic information formula which measures the uniformity of attributes domain to find DC probability. Also, FACO-Miner has some new features that make it different from existing classifiers based on ACO meta-heuristic. To classify test samples Author have defined the new fuzzy reasoning method based on averaging which takes account both the number of rules and the covering value to classify the input samples.

In Year 2011, Pooia Lalbakhsh performed a work," Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining". In this paper an improved version of Ant-Miner algorithm is introduced and compared to the previously proposed ant-based rule mining algorithms. Presented algorithm modifies the rule pruning process and introduces a dynamic pheromone evaporation strategy. The algorithm was run on five standard datasets and the average accuracy rate and numbers of discovered rules were analyzed as two important performance metrics of rule mining.

In Year 2011, Ghada Almodaifer performed a work," Discovering Medical Association Rules from Medical Datasets". In this paper, Author aim to discover interesting medical association rules from medical datasets for prediction purposes. Author provide an association rule mining system that discovers constrained association rules in medical records that includes numeric, categorical and image features.

In Year 2012, Qioling Duan performed a work," Mining Indirect Association Rules in Multi-database". Author propose an approach of synthesizing frequent itemsets before mining indirect rules in multi-database on base of previous work in multi-database mining. The experimental results demonstrate that algorithm is correct and effective.

In Year 2011, P. Kasemthaweesab performed a work," Association Analysis of Diabetes Mellitus (DM) With Complication States Based on Association Rules". Association Rule is one of important methods in data mining. By discovering data association, new useful information can be obtained. In this paper, a researcher has presented a basic method of discovering an association of diabetes mellitus with complication states by applying gender, age and occupation factors and testing to find out a relationship of diagnostic data.

In Year 2013, Divya Bhugra performed a work," Association Rule Analysis Using Biogeography Based Optimization". In this paper, Author have tried to optimize the rules generated by Association Rule Mining using Biogeography Based Optimization(BBO).BBO has a way of sharing information between solutions depending on the migration mechanisms .The motivation of this paper is to use the feature of BBO for finding more accurate results.

In Year 2013, K.Rameshkumar performed a work," Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique". This paper proposes the n-cross validation technique to reduce association rules which are irrelevant to the transaction set. The proposed approach used partition based approaches are supported to association rule validation.

HMM Model

HMM known as Hidden Markov Model. Hidden means- having set of hidden states and Markov means- next state depends only on present state. HMM is a High Level Classification. HMM is a Finite State Machine i.e. consists of set of states, start state, input, transition function, next state. HMM is a full probabilistic model. HMM helps to construct a complex model by drawing spontaneous picture. HMM is known as hidden as it work on image & generate a sequence. It is Statistical approach (set of assumptions concerning the generation of the observed data).

III. CLASSIFICATION APPROACH

CLASSIFICATION:

Classification classifies the group. Classification is the process of finding the model or function to predict the class of object whose class label is unknown. Classification is a supervised learning process. It is the machine learning technique used to predict group membership for data instances. For example of cases where the data analysis task is Classification: a loan officer in a bank want to give the loan to the customer then he wants to analyses the data in order to know which customer (loan applicant) are risky or which are safe. The above example shows the classifier is constructed to predict categorical labels.

1. Bayesian network

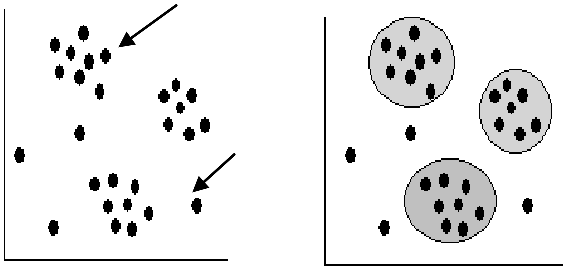
Bayesian approach is an important DM technique. Bayesian network is a decision making method. Bayesian method is based on the probability theory. This characterized the future states as probability events. Where the sum of probabilities is 1. Bayesian Networks and Probabilistic Network are known as belief network. There are two components to define Bayesian Belief Network:

- Directed acyclic graph
- A set of conditional probability tables

It represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). In DAG edges represent conditional dependencies and each node is associated with a probability function that takes as input. Bayesian network can be used to gain understanding about a problem domain and to predict the intervention. It is a statistical technique. Clustering is a unsupervised learning. Clustering is used for data analysis. Clustering algorithm is apply on similar group with similar properties for data analysis, these similar group is called cluster. Cluster therefore is a collection of objects which are similar between them and are dissimilar to object belonging to other clusters. Clustering is used for parallel processing. Clustering is an unsupervised learning. With the help of Clustering we determine the intrinsic grouping in a set of unlabeled data.

2. Clustering

Clustering is a unsupervised learning. Clustering is used for data analysis. Clustering algorithm is apply on similar group with similar properties for data analysis, these similar group is called cluster. Cluster therefore is a collection of objects which are similar between them and are dissimilar to object belonging to other clusters. Clustering is used for parallel processing. Clustering is an unsupervised learning. With the help of Clustering we determine the intrinsic grouping in a set of unlabeled data.



K mean clustering – it cluster observations into groups of related observations without any prior knowledge. K mean clustering minimize the average square distance between the points in the same cluster.One of the main disadvantages to k-means is that you must specify the input (number of clusters) to the algorithm. In k mean clustering there are always K clusters and they do not overlap. It is faster than hierarchical clustering. The aim of k mean clustering algo is to partition observations into k clusters in which each observation belongs to the cluster with the nearest mean. It is the simplest unsupervised learning algorithms

Steps - Place K points into the space. These k points represent the objects which been to clusters (target clusters). Then Assign each object to the group that has the closest centroid. Then recalculate the positions of the K centroids. Repeat until the centroids do not move. As a result of this minimized separation can be calculated between clusters.

3. Neural Network NN is a systematic step-by-step procedure to perform task. NN is AI approach. It is composed of a large number of highly interconnected processing elements working in parallel to solve specific problem.

For forecasting ability many factors can affect artificial neural networks.

1) If unequal number of neurons is used, the result will be poor.
2) If too many neurons are used, the training time may become long.
3) The selection of a learning rate is of critical importance in finding the true global minimum of the error distance.
4) Very low momentum factor causes the local minima.
5) If the lag period (gap) is smaller than required then forecasting ability will be in danger.

4. DSS (Decision Support System)

DSS is used for decision making process. Decision making activities are the steps taken to choose a suitable alternative from needed for realizing a certain goal [17]. For this information about the outcome is considered and a path is chosen which help in achieving the goal.

Decision Tree- Decision Tree is the important field of DSS. ctree function is call function to build the decision tree. It covers all the aspects of important techniques which are used to discover the pattern. Decision Tree helps to choose the place that best fits & also their need. Its structure contains of root (starting point), branch (outcome of test) and leaf node (class label). Decision Tree helps to make something known in advance therefore known as predictive model.

5. Association

Association detects sets of attributes and rules among them. It also finds the correlation of multiple databases. Association rules are if/then statements that help uncover relationships between unrelated databases. Application of Association are – clustering, classification. An association rule is about relationships between two disjoint item sets X and Y such that X ⇒ Y. It presents the pattern when X occurs, Y also occurs.

For an association rule $X \Rightarrow Y$, we can calculate

Support ($X \Rightarrow Y$) = support (XY) &

Confidence ($X \Rightarrow Y$) = support (XY)/support

Major tasks of data mining

1) Explore the data for analysis.
2) Describe all the data collected.
3) Perform inference on data to make predictions.
4) Discovering patterns and rules (detecting defects)

V. CONCLUSION

In this paper, a study on the knowledge discovery methods is provided including the clustering and classification methods. The paper has described various algorithmic methods so that the information extraction will be done effectively.

REFERENCES

[1] Adepele Olukunle," A Fast Algorithm for Mining Association Rules in Medical Image Data", Proceedings of the 2002 IEEE Canadian Conference on Electrical & Computer Engineering 0-7803-7514-9/02@2002 IEEE

[2] Carlos Ordonez," Association Rule Discovery With the Train and Test Approach for Heart Disease Prediction", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 2, APRIL 2006, 1089-7771 © 2006 IEEE

[3] Chunxue Shi," Path Planning for Deep Sea Mining Robot Based on ACO-PSO Hybrid Algorithm", 2008 International Conference on Intelligent Computation Technology and Automation

[4] Gaurav N. Pradhan," ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA", ICME 2009 978-1-4244-4291-1/09©2009 IEEE

[5] Mr.K.Ravikumar," ACO based spatial Data Mining for Traffic Risk Analysis".

[6] Wei Wang," Mining Association rules in Medical Data Based on Concept Lattice", Proceedings of the 8th World Congress on Intelligent Control and Automation July 6-9 2010, Jinan, China 978-1-4244-6712-9/10©2010 IEEE

[7] Mostafa Fathi Ganji," Parallel Fuzzy Rule Learning Using an ACO-Based Algorithm for Medical Data Mining", 978-1-4244-6439-5/10©2010 IEEE

[8] Pooia Lalbakhsh," Focusing on Rule Quality and Pheromone Evaporation to Improve ACO Rule Mining", 2011 IEEE Symposium on Computers & Informatics 978-1-61284-691-0/11©2011 IEEE

[9] Ghada Almodaifer," Discovering Medical Association Rules from Medical Datasets", 978-1- 61284-704-7/11 ©2011 IEEE

[10] Qiaoling Duan," Mining Indirect Association Rules in Multi-database", 2012 3rd International Conference on System Science, Engineering Design and Manufacturing Informatization 978-1-4673-0915-8/12©2012 IEEE

[11] P. Kasemthaweesab," Association Analysis of Diabetes Mellitus (DM) With Complication States Based on Association Rules", 978-1-4577-2119-9/12@ 2011 IEEE

[12] Divya Bhugra," Association Rule Analysis Using Biogeography Based Optimization", 2013 International Conference on Computer Communication and Informatics (ICCCI -2013), Jan. 09 – 11, 2013, Coimbatore, INDIA 978-1-4673-2907-1/13 ©2013 IEEE

[13] K.Rameshkumar," Relevant Association Rule Mining from Medical Dataset Using New Irrelevant Rule Elimination Technique".