Study of Classification Methods in Data Mining

Prof. Sandeep N. Khandare

Prof. Aniruddha Holey

Abstract: Classification is a data mining technique based on machine learning which is used to classify each item in a set of data into a set of predefined classes or groups. Classification methods make use of mathematical and statistical techniques such as decision trees, linear programming, neural network and other methods like Genetic Approach, Fuzzy set Approach, Ruled based etc. This paper is an survey of different classification methods and there advantages. In this paper Classification Method is considered, it focuses on a survey on various classification techniques that are most commonly used in data-mining.

I. INTRODUCTION

Data Mining is defined as extracting information from huge sets of data. We can also say that data mining is the procedure of mining knowledge from data.
A. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to the objectives.
B. Data Understanding: This phase starts with an initial data collection, to get familiar with the data, to identify data quality problems, to detect interesting subsets to form hypothesis for hidden information.
C. Data Preparation: This phase covers all activities to construct the final dataset from the initial raw data.
D. Modeling: This phase includes various modeling techniques are selected and applied and their parameters are calibrated to optimal values.
E. Evaluation: This phase include model is thoroughly evaluated and reviewed. At the end of this phase, a decision on the use of the data mining results should be reached.
F. Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

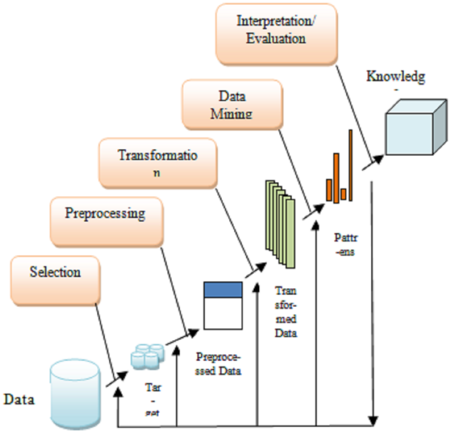

Figure:1 Data Mining Process

Classification: Classification techniques in data mining are capable of processing a large amount of data. It can predict categorical class labels and classifies data based on training set and class labels and hence can be used for classifying newly available data.

II. CLASSIFICATION MODELS

A. Decision Tree Classifier: Decision Tree Classifier (DTC) is a simple and widely used classification technique. It is a classifier in the form of a tree structure. In which there is decision node that specifies a test on a single attribute and leaf node that indicates the value of the target attribute. Arc/edge is there for split of one attribute. Path is a disjunction of test to make the final decision. It applies a straight forward idea to solve the classification problem. Decision trees classify instances or examples by starting at the root of the tree and moving through it until a leaf node. It poses a series of carefully crafted questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until a conclusion about the class label of the record is reached. 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. Decision tree classifier has limitation as it is computationally expensive because at each node, each candidate splitting field must be sorted before its best split can be found.

B) Neural Networks: Neural Networks (NNs) are models for classification and prediction. The idea behind neural networks is to combine the input information in a very flexible way that captures complicated relationships among these variables and between them and the response variable The most common action in data mining is classification. It recognizes patterns that describe the group to which an item belongs. It does this by examining existing items that already have been classified and inferring a set of rules. Similar to classification is clustering. The major difference being that no groups have been predefined. Prediction is the construction and use of a model to assess the class of an unlabeled object or to assess the value or value ranges of a given object is likely to have. The next application is forecasting. This is different from predictions because it estimates the future value of continuous variables based on patterns within the data. Neural networks, depending on the architecture, provide associations, classifications, clusters, prediction and forecasting to the data mining industry Neural networks use a set of processing elements (or nodes) analogous to neurons in the brain. These processing elements are interconnected in a network that can then identify patterns in data once it is exposed to the data, i.e., the network learns from experience just as people do. This distinguishes neural networks from traditional computing programs that simply follow instructions in a fixed sequential order. Neural networks essentially comprise three pieces: the architecture or model; the learning algorithm; and the activation functions. Neural networks are programmed or "trained to" . . . store, recognize, and associatively retrieve patterns or database entries; to solve combinatorial optimization problems; to filter noise from measurement data; to control ill defined problems. It is precisely these two abilities (pattern recognition and function estimation) which make artificial neural networks (ANN) so prevalent a utility in data mining. As data sets grow to massive sizes, the need for automated processing becomes clear. With their "model-free" estimators and their dual nature, neural networks serve data mining in a myriad of ways.

C) Genetic Programming: The idea of genetic algorithm is derived from natural evolution. In genetic algorithm, first of all, the initial population is created. This initial population consists of randomly generated rules. We can represent each rule by a string of bits.

Points to remember:

I. Based on the notion of the survival of the fittest, a new population is formed that consists of the fittest rules in the current population and offspring values of these rules as well.

1. The fitness of a rule is assessed by its classification accuracy on a set of training samples.

2. The genetic operators such as crossover and mutation are applied to create offspring.

3. In crossover, the substring from pair of rules is swapped to form a new pair of rules.

4. In mutation, randomly selected bits in a rule's string are inverted.

Three operators are used by genetic algorithms:

1. *Selection*. The selection operator refers to the method used for selecting which chromosomes will be reproducing. The fitness function evaluates each of the chromosomes (candidate solutions), and the fitter the chromosome, the more likely it will be selected to reproduce.

2. *Crossover*. The crossover operator performs recombination, creating two new offspring by randomly selecting a locus and exchanging subsequences to the left and right of that locus between two chromosomes chosen during selection. For example, in binary representation, two strings 11111111 and 00000000 could be crossed over at the sixth locus in each to generate the two new offspring 11111000 and 00000111.

3. *Mutation*. The mutation operator randomly changes the bits or digits at a particular locus in a chromosome: usually, however, with very small probability. For example, after crossover, the 11111000 child string could be mutated at locus two to become 10111000. Mutation introduces new information to the genetic pool and protects against converging too quickly to a local optimum **.**

Most genetic algorithms function by iteratively updating a collection of potential solutions called a population. Each member of the population is evaluated for fitness on each cycle. A new population then replaces the old population using the operators above, with the fittest members being chosen for reproduction or cloning.

The fitness function f (x) is a real-valued function operating on the chromosome (potential solution), not the gene, so that the x in f (x) refers to the numeric value taken by the chromosome at the time of fitness evaluation

D. Rough Set Approach : We can use the rough set approach to discover structural relationship within imprecise and noisy data.

Note: This approach can only be applied on discrete-valued attributes. Therefore, continuous-valued attributes must be discretized before its use.

The Rough Set Theory is based on the establishment of equivalence classes within the given training data. The tuples that forms the equivalence class are indiscernible. It means the samples are identical with respect to the attributes describing the data. There are some classes in the given real world data, which cannot be distinguished in terms of available attributes. We can use the rough sets to **roughly** define such classes.

E. Bayesian classification :A Bayesian network (BN) consists of a directed, acyclic graph and a probability distribution for each node in that graph given its immediate predecessors. A Bayes Network Classifier is based on a bayesian network which represents a joint probability distribution over a set of categorical attributes. It consists of two parts, the directed acyclic graph G consisting of nodes and arcs and the conditional probability tables. The nodes represent attributes whereas the arcs indicate direct dependencies. The density of the arcs in a BN is one measure of its complexity. Sparse BNs can represent simple probabilistic models (e.g., naïve Bayes models and hidden Markov models), whereas dense BNs can capture highly complex models. Thus, BNs provide a flexible method for probabilistic modeling.

F. Rule-based classification : Rule-based classifier (RBC) makes use of set of IF-THEN rules for classification. We can express the rule in the following from: IF condition

THEN conclusion. The IF part of the rule is called rule antecedent or precondition. The THEN part of the rule is called rule consequent. In the antecedent part the condition consists of one or more attribute. The consequent part consist class prediction. It is easy to interpret and generate

G. Fuzzy Set Approach: Fuzzy Set Theory is also called Possibility Theory. This theory was proposed by Lotfi Zadeh in 1965 as an alternative the **two-value logic** and **probability theory**.

This theory allows us to work at a high level of abstraction. It also provides us the means for dealing with imprecise measurement of data.

Fuzzy logic is a form of many-valued logic; it deals with reasoning that is approximate rather than fixed and exact. Compared to traditional binary sets. Fuzzy logic variables may have a truth value that ranges in degree between 0 and 1. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false.

In mathematical logic, there are several formal systems of "fuzzy logic"; most of them belong among so-called t-norm fuzzy logics.

Propositional fuzzy logics:

1.Monoidal t-norm-based propositional fuzzy logic MTL is an axiomatization of logic where conjunction is defined by a left continuous t-norm, and implication is defined as the residuum of the t-norm. Its models correspond to MTL-algebras that are prelinear commutative bounded integral residuated lattices.

2. Basic propositional fuzzy logic BL is an extension of MTL logic where conjunction is defined by a continuous t-norm, and implication is also defined as the residuum of the t-norm. Its models correspond to BLalgebras.

3. Łukasiewicz fuzzy logic is the extension of basic fuzzy logic BL where standard conjunction is the Łukasiewicz t-norm. It has the axioms of basic fuzzy logic plus an axiom of double negation, and its models correspond to MV-algebras.

4. Gödel fuzzy logic is the extension of basic fuzzy logic BL where conjunction is Gödel t-norm. It has the axioms of BL plus an axiom of idempotence of conjunction, and its models are called G-algebras.

5.Product fuzzy logic is the extension of basic fuzzy logic BL where conjunction is product t-norm. It has the axioms of BL plus another axiom for cancel activity of conjunction, and its models are called product algebras.

6. Fuzzy logic with evaluated syntax (sometimes also called Pavelka's logic), denoted by EVŁ, is a further generalization of mathematical fuzzy logic. While the above kinds of fuzzy logic have traditional syntax and many-valued semantics, in EVŁ is evaluated also syntax. This means that each formula has an evaluation. Axiomatization of EVŁ stems from Łukasziewicz fuzzy logic. A generalization of classical Gödel completeness theorem is provable in EVŁ

H) Ant Colony: Ant Colony algorithms were first introduced in the 1992 PhD thesis of Marco Dorigo. They offer a way of finding good paths within a graph, and they were inspired by the behavior of ants in finding paths from the colony to food. When traveling from their colony to food sources, ants deposit chemicals called pheromones on the trails. The trail is used so that ants can find their way back to the colony. If other ants find this path they are likely to follow it. This will cause more pheromone to be deposited on the trail, which has the effect of reinforcing it. However, if a trail has not been used for a while, the pheromone starts to Evaporate. Short paths have the advantage of being marched over faster and more often, therefore, the pheromone density remains high. So a short path will have more ants traveling on it, increasing the pheromone density even more, and eventually all the ants will follow it. Ant algorithms mimic this behaviour in order to find optimal paths within a graph. Initially, all paths have a random small amount of pheromone deposited on it. An ant departs from the starting node and starts the process of visiting the other nodes in the graph. At each node, the ant decides which node it should visit next. The ant rates the attractiveness of traveling to each node in the graph. A nearby node with a large amount of pheromone deposited in its path will be very attractive. Also, a distant node with little amount of pheromone deposited in its path will be very unattractive. This attractiveness information will be use in order to compute the probability that a given route will be taken. Ant algorithms give good results relatively fast. They can be run continuously in order to adapt to changes in real-time: an advantage as compared to genetic algorithms. For instance, an obstacle such as a traffic jam would quickly lead the ants to discover a different good path. This can be of interest in applications such as network routing and urban transportation.

Analysis of Classification methods is tabulated in Table 1.

Table:1

Methods	Algorithms	Advantages	Author
Decision Tree	Decision Tree Induction Algorithm (ID3,C4.5)	1.It does not require any domain knowledge. 2It is easy to comprehend. 3 It is simple and fast.	J. Ross Quinlan
Neural Networks	The Back Propagation Algorithm	1. High Accuracy 2. Noise Tolerance 3. Independence from prior assumptions 4. Ease of maintenance. 5. can be implemented in	

		parallel hardware	
Bayesian classification	Bayes' Theorem	1.To improves the classification performance by removing the irrelevant features. 2.Good performance. 3.it is short computational time	Thomas Bayes.
Genetic	Genetic algorithm		John Holland
Rule-based	Sequential Covering Algorithm		
Fuzzy Set Approach		Allows working at a high level of abstraction. It provides the way to deal with mprecise measurement of data.	Lotfi Zadeh
Ant Colony	*Ant colony Algorithm*		Marco Dorigo

CONCLUSION

Data mining is a broad area that integrates techniques from several fields including machine learning, statistics, pattern recognition, artificial intelligence, and database systems, for the analysis of large volumes of data. There have been a large number of data mining algorithms rooted in these fields to perform different data analysis tasks In this paper we studied commonly use approaches of classification performance and advantages .

REFERENCES

[1] N.Abirami1,T.Kamalakannan 2,Dr.A. Muthukumaravel3" A Study on Analysis of Various Datamining Classification Techniques on Healthcare Data"

[2]Dr. Yashpal Singh,Alok Singh Chauhan"Neural Networks in Data Mining

[3] Hem Jyotsana Parashar, Singh Vijendra, and Nisha Vasudeva"An Efficient Classification Approach for Data Mining"

[4]Delveen Luqman Abd AL-Nabi1* Shereen Shukri Ahmed2 "Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation)"

[5]S.Archana1, Dr. K.Elangovan2 "Survey of Classification Techniques in Data Mining"

AUTHOR'S PROFILE

Prof. Sandeep N. Khandare
Asst. Professor
MCA ,
P.G. Department of Computer Science and Tech, DCPE,HVPM, Amravati

Prof. Aniruddha Holey
Asst. Professor
MCA ,
Member of Society for Promotion of Excellence in Electronics Discipline (SPEED)
P.G. Department of Computer Science and Tech, DCPE,HVPM, Amravati