A Survey of Classification Methods and its Applications

Geetika ,
PhD Scholar,
ITM University, Gurgaon

ABSTRACT

Data classification is the categorization of data for its most effective and efficient use. Data can be classified according to any criteria, not only relative importance or frequency of use. Classification can help an organization to meet legal and regulatory requirements for retrieving specific information within a set timeframe, and this is often the motivation behind implementing data classification methods and algorithms. The paper contains brief discussion of various classification methods that includes decision trees, K-nearest neighbor classifier, naïve bayes classifier and neural network. The paper also discusses some applications of classification model and at the end the paper is concluded with the brief observations of these classification models.

1. INTRODUCTION

Data Mining is the nontrivial process of identifying valid, novel, potentially useful and ultimately understandable pattern in data with the wide use of databases and the explosive growth in their sizes. Data mining refers to extracting or “mining” knowledge from large amounts of data. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data. Data mining usually involves four tasks: Classification, Clustering, Association Rule Learning, and Regression. The essential process of Knowledge Discovery is the conversion of data into knowledge in order to aid in decision making, referred to as data mining[1]. Knowledge Discovery process consists of an iterative sequence of data cleaning, data integration, data selection, data mining pattern recognition and knowledge presentation. Data mining is the search for the relationships and global patterns that exist in large databases but are hidden among large amounts of data.

2. CLASSIFICATION AND REGRESSION

Regression models map the input space into a real-valued domain. For instance, a regressor can predict the demand for a certain product given its characteristics. On the other hand, classifiers map the input space into redefined classes. For instance, classifiers can be used to classify mortgage consumers as good (full mortgage pay back the on time) and bad (delayed pay back). Among the many alternatives for representing classifiers, there are, for example, support vector machines, decision trees, probabilistic summaries which are discussed in coming sections. Classification routines in data mining also use a variety of algorithms and the particular algorithm used can affect the way records are classified.

2.1 Decision tree Classifiers

A common approach for classifiers is to use decision trees to partition and segment records. New records can be classified by traversing the tree from the root through branches and nodes, to a leaf representing a class. The path a record takes through a decision tree can then be represented as a rule. For example, if "Income<=102000 and age<58, then individual is high spender (figure 1). But due to the sequential nature of the way *a* decision tree splits records (i.e. the most discriminative attribute-values [e.g. Income] appear early in the tree) can result in a decision tree being overly sensitive to initial splits. Therefore, in evaluating the goodness of fit of a tree, it is important to examine the error rate for each leaf node (proportion of records incorrectly classified). A nice property of decision tree classifiers is that because paths can be expressed as rules, then it becomes possible to use measures for evaluating the usefulness of rules such as Support, Confidence.

In data mining, a decision tree is a predictive model which can be used to represent both classifiers and regression models. When a decision tree is used for classification tasks, it is more appropriately referred to as a classification tree. When it is used for regression tasks, it is called regression tree. Tree construction proceeds recursively starting with the entire set of training examples. At each step, an attribute is selected as the root of the (sub) tree and the current training set is split into subsets according to the values of the selected attribute. For discrete attributes, a branch of the tree is typically created for each possible value of the attribute. For continuous attributes, a threshold is selected and two branches are created based on that threshold. For the subsets of training examples in each branch, the tree construction algorithm is called recursively. Tree construction stops when the examples in a node are sufficiently pure (i.e., all are of the same class) or if some other stopping criterion is satisfied (there is no good attribute to add at that point). Such nodes are called leaves and are labeled with the corresponding values of the tree. The finding of a solution with the help of decision trees starts by preparing a set of solved cases.

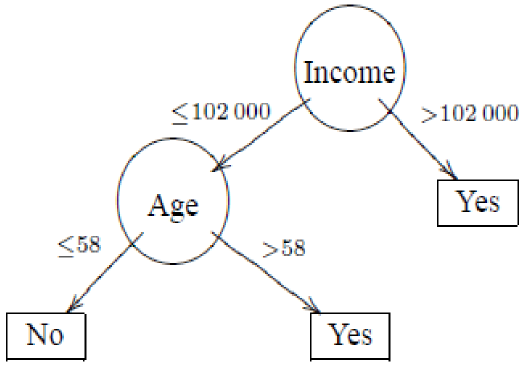

Figure 1 Decision tree example for High spender

2.1.1 Decision Tree construction algorithm

ID3, C4.5 and CART some of tree construction algorithms, ID3 and C4.5 works on recursive partitioning. Their basic idea is to partition the sample space in a data-driven manner, and represent the partition as a tree. An important property of these algorithms is that they attempt to minimize the size of the tree at the same time they optimize it, some quality measure features of a new sample are matched against the conditions of the tree. Hopefully, there will be exactly one leaf node whose conditions (on the path) will be satisfied. Decision tree algorithms can select the best features from a given set, they generally cannot build new features needed for a particular task, and another important feature of ID3 trees is that each attribute can provide at most one condition on a given path. This also contributes to comprehensibility of the resulting knowledge. ID3 uses chi square test (statistical calculation) in addition to information gain from a single attribute to build a decision tree. i.e. information gain given by the formula *gain (X) = info (T) - infoX (T).* Where *info* is actually entropy, *info (T) is overall entropy and infoX(T) is entropy of X attribute*. Information gain is biased towards tests with many outcomes. Gain ratio used in c4.5 criterion (Quinlan, 1993) was developed to avoid this bias. The information generated by dividing T into n subsets is given by split info which is further used to produce *gain ratio* (*X*) = *gain* (*X*) */ split info* (*X*) [2].

Classification and Regression Tree In CART the conditions in the tree are based on thresholds for continuous domain Because of that, the conditions on a path can use a given attribute a number of times (with different thresholds), and the thresholds used on different paths are very likely to differ, It can handle high dimensional categorical data.. Also, CART is capable of inducing new features, restricted to linear combinations of the existing features in regression trees; the leaves predict a real number and not a class. In case of regression, CART looks for splits that minimize the prediction squared error (the least-squared deviation). The prediction in each leaf is based on the weighted mean for node.

One of the most significant advantages of decision trees is the fact that knowledge can be extracted and represented in the form of classification (if-then) rules. Each rule represents a unique path from the root to each leaf. First, all attributes defining each case are described (input data) and among them one attribute is selected that represents a decision for the given problem (output data). For all input attributes specific value classes are defined. If an attribute can take only one of a few discrete values then each value takes its own class; if an attribute can take various numeric values then some characteristic intervals must be defined, which represent different classes. Each attribute can represent one internal node in a generated decision tree, also called an attribute node or a test node. Such an attribute node has exactly as many branches as its number of different value classes. The leaves of a decision tree are decisions and represent the value classes of the decision attribute – decision classes. When a decision has to be made for an unsolved case, we start with the root node of the decision tree and moving along attribute nodes select branches where values of the appropriate attributes in the unsolved case matches the attribute values in the decision tree until the leaf node is reached representing the decision.

Decision trees are recognized as highly unstable classifiers with respect to minor perturbations in the training data, in other words, methods presenting high variance. Fuzzy logic brings in an improvement in these aspects due to the elasticity of fuzzy sets formalism. The satisfaction level of each rule must be determined and a conflict resolution used. one can determine how a data input for a fuzzy variable, satisfies the fuzzy restriction we can also combine levels of satisfaction of fuzzy restrictions of the conjunctive antecedent the resulting value is often called degree of fulfillment or satisfaction of the rule.

2.2. Naive Bayes Classifier network

It provides new ways of exploring and understanding data. It learns from the “evidence” by calculating the correlation between the target i.e. dependent and other i.e., independent variables. The probabilistic model of NBC is to find the probability of a certain class given multiple disjoint (assumed) events. The Naïve Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f(x) can take on any value from some finite set V. A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values <a1, a2, an>. The learner is asked to predict the target value, or classification, for this new instance. A Bayesian network is a representation of the joint distribution over all the variables represented by nodes in the graph. Let the variables be X(1), ..., X(n). Let parents (A) be the parents of the node A, then the joint distribution for X(1) through X(n) is represented as the product of the probability distributions P(Xi|Parents(Xi)) for i = 1 to n. If X has no parents, its probability distribution is said to be unconditional, otherwise it is conditional. The conditional probability values of all the attributes with respect to the class are pre-computed and stored on disk. This prevents the classifier from computing the conditional probabilities every time it runs[4].

This stored data can be reused to reduce the latency of the classifier. The most interesting feature of Bayesian Networks, compared to decision trees is most certainly the possibility of taking into account prior information about a given problem. An important advantage of Naive Bayes is that the simple structure lends itself to comprehensible visualizations. Bayesian networks can readily handle incomplete data sets. Bayesian networks allow one to learn about causal relationships Bayesian networks readily facilitate use of prior knowledge. Bayesian classifiers are used when the data is high, the attributes are independent of each other and when we want more efficient output, as compared to other methods output.

2.3. K-Nearest Neighbor classifier

KNN assumes that the data is in a feature space. More exactly, the data points are in a metric space. The data can be scalars or possibly even multidimensional vectors. Since the points are in feature space, they have a notion of distance this need not necessarily be Euclidean distance although it is the one commonly used. Each of the training data consists of a set of vectors and class label associated with each vector. In the simplest case, it will be either + or – (for positive or negative classes). But KNN, can work equally well with arbitrary number of classes. We are also given a single number "k". This number decides how many neighbors (where neighbors are defined based on the distance metric) influence the classification. This is usually an odd number if the number of classes is 2. If k=1, then the algorithm is simply called the nearest neighbor algorithm. In this algorithm, we are given some data points for training and also a new unlabelled data for testing. Our aim is to find the class label for the new point. Larger k values help reduce the effects of noisy points within the training data set, and the choice of k is often performed through cross-validation. In KNN, k is usually chosen as an odd number if the number of classes is 2. Choice of k is very critical, a small value of k means that noise will have a higher influence on the result. A large value makes it computationally expensive and kind of defeats the basic philosophy behind KNN (that points that are near might have similar densities or classes).

2.4. Neural network classifiers

The classifiers generated by neural networks are described as complex mathematical functions; they are rather incomprehensible and opaque to humans. They follow discriminating rule to classify data e.g. in a in a two-group classification problem, if the desired output is coded as 1 if the object is from class 1 and if it is from class 2. NN opacity limits them in many real-life applications where both accuracy and comprehensibility are required, such as medical diagnosis and credit risk evaluation [5].

Application of classification models

3. APPLICATIONS OF CLASSIFIERS

3.1 Detecting adverse drug events

Data are extracted from several hospitals' electronic health records (EHRs) to feed a common repository with past fully anonymized hospital stays. The meaning full attributes for classification purpose person's age, gender, admission date, laboratory results, drugs taken, doctor's prescription. Aggregation engine will be used to transform the available data in to information described as set of events so that data can be converted into binary form. Aggregation example there are several measures of potassium available in patients data, its level in the blood should not reach too low or too high values; otherwise, it could lead to lethal heart arrhythmias[6]. The repeated measures are taken from 2-6 days the data taken is the converted in to binary value high as 1 and low as 0. The knowledge about ADEs can be expressed using rules where some conditions lead to an outcome. Some of the variables computed by the aggregation process can be used as outcome (e.g., death) and some other ones can be used as conditions (e.g., chronic renal failure). Each condition must be an event that occurs before the outcome and is still active or has ended less than a fixed delay before the outcome occurs[6].

3.2 Credit risk evaluation

Individual credit risk evaluation is an important and challenging data mining problem in financial analysis domain[3]**.** Credit risk is referred to as the risk of loss when a debtor does not fulfill its debt contract and is of natural interest to practitioners in bank as well as to regulators. The credit problem can also become a crisis when some of the risk lands back on banks. Each sample has 20 attributes (7 numerical and 13 categorical), including account status, loan purposes, loan amount, age, property status, etc.

3.3 Heart disease prediction

The diagnosis of diseases is a significant and tedious task in medicine. The detection of heart disease from various factors or symptoms is a multi-layered issue which is not free from false presumptions often accompanied by unpredictable effects. Users enter values of medical attributes to diagnose patients with heart disease. For example, attributes such as age, sex, chest pain, blood sugar, blood pressure can predict that this patient has a heart disease[4]. When the significant attributes have high values, doctors could recommend that the patient should undergo further heart examination. Thus performing "what if" scenarios can help prevent a potential heart attack [4]. Doctors can use this information to further analyze the strengths and weaknesses of the medical attributes associated with heart disease. Identify the impact and relationship between the medical attributes in relation to the predictable state heart disease. Identifying the impact and relationship between the medical attributes in relation to heart disease is only found in Decision Trees viewer. Doctors can use this information to perform medical screening on selected significant attributes instead of on all attributes on patients who are likely to be diagnosed with heart disease. This will reduce medical expenses, administrative costs, and diagnosis time.

4. CONCLUSION AND OBSERVATIONS

Decision trees and rule classifiers have a similar operational profile. The goal of classification result integration algorithms is to generate more certain, precise and accurate system results. Fuzzy representation of trees add flexibility in data, the fuzzy decision tree can produce real valued outputs with gradual shifts. Moreover, fuzzy sets and approximate reasoning allow for processing of noisy and inconsistent/incomplete data. The results obtained in [6] include 236 rules the fuzzy approach can add satisfaction to these rules and may omit or increase some rules as fuzzy sets represent more flexibility in data

The nearest-neighbor method suffers severely from what is called the "curse of dimensionality."More subtly, the accuracy of the method tends to deteriorate as m increases. In decision tree small variations in the training set, the algorithm may choose an attribute which is not truly the best one. Naïve Bayes appears to fare better than Decision Trees as it shows the significance of all input attributes incase of heart disease prediction. The regression model is generally used in medical diagnosis. The observations are summarized in table 1.

Classifier model	Observations
C4.5	Sensitive to input data, Good at categorical data. Accuracy can be increased by feature selection. Rules slow for large and noisy datasets.
Bayesian network	Work well only when underline assumption are satisfied. Good knowledge of data and model capabilities are necessary
Neural network	Power full data fitting or function approximation makes it susceptible to over fitting problem. Combining several neural networks can be used to improve the performance.

Table 1 summary of classification models

5. REFERENCES

[1]. Han J., Kamber M. "*Datamining Concepts and Techniques*" Second edition, Elsevier, PP 285-295.

[2]. Vili P, Peter K, Bruno S, Ivan R. **"***Decision trees: an overview and their use in medicine***"** Journal of Medical Systems Vol. 26, Num. 5, pp. 445-463, October 2002.

[3]. Hong Yu, Xiaolei Huang, Xiaorong Hu, Hengwen Cai "*A Comparative Study on Data Mining Algorithms for Individual Credit Risk Evaluation*" IEEE conference 2012.

[4] G.Subbalakshmi, K. Ramesh, Chinna Rao *Decision Support in Heart Disease Prediction System using Naive Bayes"* G.Subbalakshmi, K. Ramesh, Chinna Rao,IJCSE 2011.

[5] Geetika M , Sunint K."*Comparative study of ANN for pattern classification*" WSEAS Int. Conf. on Mathematical Methods and Computational Techniques in Electrical Engineering, Bucharest, October 2006.

[6] Emmanuel C, Gregoire F., Stephanie B., Michel L., and Regis B."*Data Mining to Generate Adverse Drug Events Detection Rules"* IEEE Transactions on Information technology in Biomedicine,Vol 15, No. 6 Nov 2011.