ISSN No. 0976-5697

Volume 3, No. 5, Sept-Oct 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Decision Support through Hierarchical Multi-Attribute Predictive models

Pankaj Pathak*
PhD Scholar (Comp Sc.)
Pacific Academy of Higher Education and Research
University Udaipur, India
pankajpathak101@gmail.com

Dr. P. R. Pal
Professor,
Department of Computer Applications
Ajay Kumar Garg Engineering College
Ghaziabad, India
prpal@rediffmail.com

Abstract: In present scenarios every industry is facing a critical competition. As the number of choices are available to the customers of every type of product. Most of the industries are runs their operations in several countries. So each industry wants to increase their valuable customers by knowing the buying behavior and the preferences of the customers which are based upon number of factors. To derive this knowledge about the customers with the help of existing data sets and information is possible by Data Mining Techniques. We can make important strategic decisions on the basis of results of these data mining techniques. One of the important data mining techniques is decision tree a hierarchical structure which contains nodes and directed edges. The Predictive models can build with the help of Decision Trees.

Keywords: Decision, Data Mining, Decision Tree, Classification, Uncertainty

I. INTRODUCTION

Hierarchical multi-attribute model of Data Mining Technique is powerful for classification and prediction. Classification, which is a predictive task, looks at assigning objects to one of several predefined categories or class values. It is used for facilitate the decision making process in sequential decision problems. It is a graphical model which is used to describe decisions and their outcomes those are possible. A decision tree classifies various items of data sets by applying a number of questions which are all about the features associated with data items. Each node of the decision tree contains a question and each internal node for every possible answer of its question point to one child node. Every time it receives an answer, a follow-up question is asked until it gets the conclusion about the record's class label.

In many applications, however, data uncertainty is common. The value of a attribute is not captured by a single point value, but by a range of values giving rise to a probability distribution. Data uncertainty arises in many applications due to various reasons like measurement errors, data missing, repeated measurements, limitations of the data collection process, etc. A simple way to handle data uncertainty is to abstract probability distributions by summary statistics such as means and variances.

II. BUILDING THE DECISION TREE

The decision making process is complex because of uncertainty associated with it. there are often a large number of different factors that must be taken into account when making the decision, like it may be useful to consider [1, 8] the possibility of reducing the uncertainty in the decision by collecting additional information, and the decision maker's attitude toward risk taking can impact the relative desirability of different alternatives.

In Decision Trees the set of possible input values are taken. These values perform feature selection to identify the attributes and values that provide the most information, and remove from consideration the values that are very rare. Decision tree model also create groupings of values that can be processed as a unit to optimize performance. A Decision tree structure having a single parent node which contains metadata. Under the parent node there are separate independent trees which represent the predictable attributes that is to be select. For example, if you set up your decision tree model to predict whether customers will purchase something, and provide inputs for gender and income, the model would create a single tree for the purchasing attribute, with many branches that divide on conditions related to gender and income.

III. HOW A DECISION TREE WORKS

Information produced by data mining techniques can be represented in many different ways. Decision tree structures are a common way to organize classification schemes. In classifying tasks, decision trees visualize what steps are taken to arrive at a classification. Every decision tree begins with what is termed a root node, considered to be the "parent" of every other node. Each node in the tree evaluates an attribute in the data and determines which path it should follow. Typically, the decision test is based on comparing a value against some constant. Classification using a decision tree [2, 9] is performed by routing from the root node until arriving at a leaf node. Decision trees can represent diverse types of data. The simplest and most familiar is numerical data. It is often desirable to organize nominal data as well. Nominal quantities are formally described by a discrete set of symbols. For example, weather can be described in either numeric or nominal fashion. We can quantify the temperature by saying that it is 11 degrees Celsius or 52 degrees Fahrenheit. We could also say that it is cold, cool,

mild, warm or hot. The former is an example of numeric data, and the latter is a type of nominal data. More accurately, the example of cold, cool, mild, warm and hot is a special type of nominal data, described as ordinal data. Ordinal data has an implicit assumption of ordered relationships between the values. Continuing with the weather example, we could also have a purely nominal description like sunny, overcast and rainy. These values have no relationships or distance measures.

The type of data organized by a tree is important for understanding how the tree works at the node level. Recalling that each node is effectively a test, numeric data is often evaluated in terms of simple mathematical inequality. For example, numeric weather data could be tested by finding if it is greater than 10 degrees Fahrenheit. Nominal data is tested in Boolean fashion; in other words, whether or not it has a particular value. The illustration shows both types of tests. In the weather example, outlook is a nominal data type. The test simply asks which attribute value is represented and routes accordingly. The humidity node reflects numeric tests, with an inequality of less than or equal to 70, or greater than 70.

Classifying a test record is straightforward once a decision tree has been constructed. Starting from the root node we apply the test condition to the record and follow the appropriate branch based on the outcome of the test. This will lead us either to another internal node for which a new test condition is applied or the leaf node. The class label associated with the leaf node is then assigned to the record.

IV. CLASSIFICATION

To classify the population of records and to develop a model, classification is the most commonly applied data mining technique, which employs a set [2] of pre-classified examples.

Fraud detection and credit risk applications are particularly well suited to this type of analysis. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data sets. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. To develop a classifier model classifier-training algorithm uses the pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then describes these parameters into a model called a classifier.

A. Types of classification models:

a. Classification by decision tree induction
b. Bayesian Classification
c. Neural Networks
d. Support Vector Machines (SVM)
e. Classification Based on Associations

V. DECISION TREE CLASSIFICATION

Decision trees are most widely used in data mining for the purpose of classifying data to extract useful information. It has been applied in several contexts such as medical diagnosis, signal and image pattern recognition, and demographic studies in transportation planning. The formation of decision trees is a two-phase process: a *tree-growing phase* followed by a *pruning procedure* [3]. In the tree-growing phase, a training data set is successively partitioned into nodes according to a classification scheme based on a set of designated *independent* variables that represent certain attributes of interest. (For example, in a medical diagnosis context where the data points represent age, gender and symptoms etc. On the basis of these attributes system gives medical prescription. In fig:1 if the age of patient is <20 then Prescription P0 is suggested and if age is greater than 20 gender is checked and accordingly Test will performed and different prescriptions are suggested).

At each successive step in this process, a leaf node of the tree is selected and is partitioned into two sub nodes based on some identified independent variable according to the criteria of values could be lesser or greater than a specified value. The objective of this process is to construct a tree that classifies [4] within each of its leaf nodes a set of data points that bear some common similarities with respect to certain *dependent* variables that measure relevant performance characteristics. The process continues until all the leaf nodes are refined, and they are declared to be *terminal nodes.*

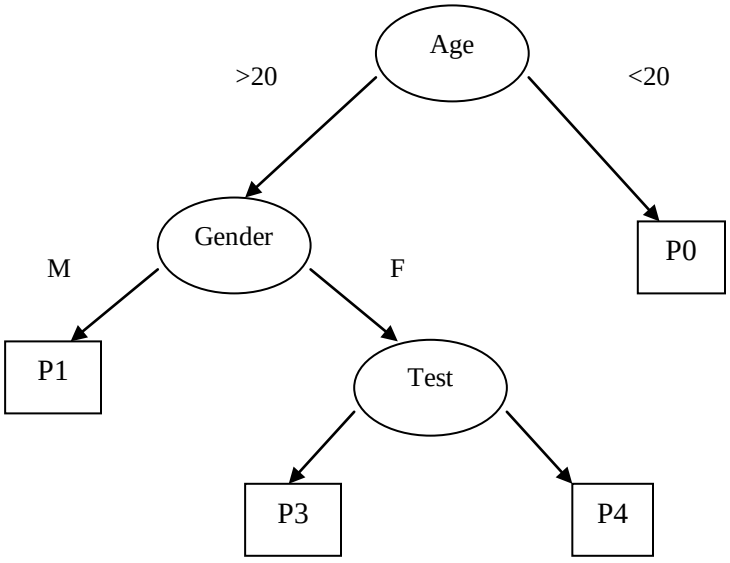

Figure 1. Decision Tree for Decision Support in medical field

VI. DECISION TREE WITH UNCERTAIN DATA

Traditional decision tree classifiers [5] work with data whose values are known and precise. For any research while searching the data sets appropriate for our Experiments, we have encountered a big obstacle is there are few data sets with complete uncertainty information. And many data sets

with numerical attributes have been collected with repeated measurements; very often the collected data have already been processed and replaced by aggregate values, such as the mean. Decision tree classification on uncertain data has been referred in the form of missing values [1], [2] and mean values. Missing values appear when some attribute values are not available during the collection of information or due to data entry errors. The Solution of missing value is to approximate these values with either by exact or probabilistic values using a classifier on the attribute (e.g., ordered attribute tree [1] and probabilistic attribute tree). Missing values in training data are handled by using fractional data. While testing, missing value is replaced by multiple values with probabilities based on the training data, thus, it allows probabilistic classification results. Extensive experiments have been conducted which show that the resulting classifiers are more accurate than those using value averages. Since processing pdfs is computationally more costly than processing single values (e.g., averages), decision tree construction on uncertain data is more CPU demanding than that for certain data.

VII. PRACTICAL APPLICATIONS

In this section, some recent successes in applying decision tree learning to solve [6, 7, 8, 9] real-world problems.

A. Predicting Library Book Use:

In [6, 7], decision trees are developed that predict the future use of books in a library. Forecasting book usage helps librarians to select less frequently used titles and move them to relatively distant and less expensive off-site locations that use efficient compact storage techniques. For this task, it is important to adopt a book choice strategy that minimizes the expected frequency of requesting removed titles. For any choice policy, this frequency depends on the percentage of titles that have to be removed for off-site storage the higher this percentage is, the higher this frequency is expected to be.

B. Intrusion Detection:

There is an increasing demand for techniques to detect intrusions into a computer and network system for information security and assurance. One of the applications of a data mining technique, decision trees, to automatically learn and recognize intrusion signatures for intrusion detection. In our study, decision tree classifiers are used to classify activities in a computer and network system into different states and determine the possibility of an intrusion based on the state classification.

C. Machine Learning:

Machine learning is another application of decision tree. It is used to perform accurately on new, unseen examples after training on a finite data set. The core objective of a learner is to generalize from its experience. The training examples from its experience come from some generally unknown probability distribution and the learner has to extract from them something more general, something about that distribution that allows it to produce useful answers in new cases.

D. Diagnosis:

As a subfield in artificial intelligence, Diagnosis is concerned with the decision making and techniques that are able to determine whether the behavior of a system is correct or not. If the system is not functioning correctly, the Decision model should be able to determine, as accurately as possible, which part of the system is failing, and which kind of fault it is facing. The computation is based on *observations*, which provide information on the current behavior.

The expression *diagnosis* also refers to the answer of the question of whether the system is malfunctioning or not, and to the process of computing the answer. This word comes from the medical context where a diagnosis is the process of identifying a disease by its symptoms.

E. Choosing the right multicore software architecture:

A multicore “decision tree” to select that multicore software architecture best suited to the application space under consideration. The first decision is to decide whether the programming model should be Symmetric Multiprocessing (SMP) or Asymmetric Multiprocessing (AMP), keeping in mind that the application can be partitioned to support both.

Choose SMP if one operating system will be run, using all of the cores as equal processing resources, and the applications can be parallelized to benefit from SMP systems. SMP requires application analysis to identify opportunities for parallelism in the code and then rewriting the code to achieve this parallelism using multithreading. For CPU intensive code, which is difficult to redesign for parallel processing using SMP and multithreading, asymmetric multiprocessing (AMP) could be a good alternative solution.

VIII. CONCLUSION

In this paper, we discuses about Decision tree, how it can build and classify the attributes to build a predictive model for decision support. It also demonstrates the decision tree with uncertain data. It is possible to achieve satisfactory classification and prediction accuracy even when data is highly uncertain. Finally the real world applications of decision tree are explained.

IX. REFERENCES

[1]. J. Dyche, the CRM Handbook: A Business Guide to Customer Relationship Management. Addison-Wesley, 2001. Pages 63-82.

[2]. Introduction to Data Mining by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Pearson Education India pages-150-175

[3]. R. Cheng, D. V. Kalashnikov, and S. Prabhakar, "Querying imprecise data in moving object Environments," IEEE Trans. Knowl. Data Eng., vol., no. 9, pp. 1112–1127, 2004

[4]. W. K. Ngai, B. Kao, C. K. Chui, R. Cheng, M. Chau, And K. Y. Yip, "Efficient clustering of uncertain Data," in ICDM. Hong Kong, China: IEEE Computer Society, 18–22 Dec. 2006, pp. 436–445.

[5]. X.Qin, Y. Zhang, X. Li, and Y. Wang, "Associative classifier for uncertain data," in international conference on Web-age information management (WAIM), pp. 692-703, 2010.

[6]. http://www.eetimes.com/design/embedded/4372692/A-decision-tree-approach-to-picking-the-right-embedded-multicore-software-architecture

[7]. Aggarwal C (2007) on density based transforms for uncertain data mining. In ICDE, pp. 866-875.

[8]. X.Qin, Y. Zhang, X. Li, and Y. Wang, "Associative classifier for uncertain data," in international conference on Web-age information management (WAIM), pp. 692{703, 2010.

[9]. C. X. Ling and C. Li. Data mining for direct marketing – specific problems and solutions. In Proceedings of Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), pages 73 – 79. 1998.