IJCSC Volume 4 • Number 1 • January-June 2013 pp. 86-88 ISSN: 0973-7391

Analysis of Decision Tree Technique of Data Mining

Sukhdev Singh[1] , Arun Jain[2], Anish Soni[3]

[1] Department of Computer Science, H.I.E.T. Kaithal, Haryana, India

[2&3] Department of CS, HCTM, Kaithal, Haryana, India-136027

sukhdev_kuk@rediffmail.com, erarunjain@rediff.com , soni_anish@yahoo.com

ABSTRACT

Advances in technology have enabled us to collect data from observations, simulations and experiments at an ever- increasing pace. For the scientist to benefit from these enhanced data collecting capabilities, it is becoming clear that semi-automated techniques, such as the ones in data mining, must be applied to find the useful information in the data. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. Decision tree is one common technique used in data mining to extract predicted information. Due to its inherent parallelism, it has been widely adopted in high performance applications and developed into various parallel decision tree algorithms in order to deal with huge datasets and complex computations.

Key words— Decision Tree, Clustring, Neural Network, Rule Induction.

I. Introduction

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Consequently, data mining consists of more than collecting and managing data; it also includes analysis and prediction.

It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Data mining can be performed on data represented in quantitative, textual, or multimedia forms. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is a relatively unique process. Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data sets. These tools can include statistical models, mathematical algorithms, and machine learning methods (algorithms that improve their performance automatically through experience, such as neural networks or decision trees).

In most standard database operations, nearly all of the results presented to the user are something that they knew existed in the database already. A report showing the breakdown of sales by product line and region is straightforward for the user to understand because they intuitively know that this kind of information already exists in the database. If the company sells different products in different regions of the country, there is no problem translating a display of this information into a relevant understanding of the business process.

Data mining, on the other hand, extracts information from a database that the user did not know existed. Relationships between variables and customer behaviors that are non-intuitive are the jewels that data mining hopes to figure out. And because the user does not know beforehand what the data mining process has discovered, it is a much bigger leap to take the output of the system and translate it into a solution to a business problem.

Data mining is used for a variety of purposes in both the private and public sectors such as Data mining is applicable in Financial sectors, health centers, oil and gas industries. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales

II. Data mining techniques

Data Mining techniques are broadly divided into *Classical Techniques*: Statistics, Neighborhoods and *Clustering Next Generation Techniques*: Decision Trees, Neural Networks and Induction Rules

These two sections have been broken up based on when the data mining technique was developed and when it became technically mature enough to be used for business, especially for aiding in the optimization of customer relationship management systems. Description of the techniques are given as follows:

a)Statistics- statistical techniques are driven by the data and are used to discover patterns and build predictive models.

b) Neighborhood- Nearest neighbor is a prediction technique that is quite similar to clustering

c) Clustering- Clustering is the method by which like records are grouped together. Clustering and the Nearest Neighbor prediction technique are among the oldest techniques used in data mining.

d) Decision Trees- A decision tree (or diagram) is a decision support tool that uses a graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

e) Neural Networks- Neural networks are biological systems that detect patterns, make predictions and learn. The artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases.

f) Rules Induction- The rules that are pulled from the database are extracted and ordered to be presented to the user based on the percentage of times that they are correct and how often they apply. In this paper we are focusing on the latest technique for data mining that is Decision tree.

III. DECISION TREE DATA MINING TECHNIQUE

A Decision Tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification. The Decision Tree is one of the most popular classification algorithms in current use in Data Mining and Machine Learning. Following are some of the features of Decision Tree.

- It divides up the data on each branch point without losing any of the data.
- The data conserved as you move up or down the tree .
- It is pretty easy to understand how the model is being built (in contrast to the models from neural networks or from standard statistics).

From a business perspective decision trees can be viewed as creating a segmentation of the original dataset (each segment would be one of the leaves of the tree). In the past this segmentation has been performed in order to get a high level view of a large amount of data - with no particular reason for creating the segmentation except that the records within each segmentation were somewhat similar to each other.

In this case the segmentation is done for a particular reason - namely for the prediction of some important piece of information. The records that fall within each segment fall there because they have similarity with respect to the information being predicted - not just that they are similar - without similarity being well defined. Thus the decision trees and the algorithms that create them may be complex, the results can be presented in an easy to understand way that can be quite useful to the business user.

IV. DECISION TREE ANALYSIS PROCESS

g) Step1*: Data Exploration:* The decision tree technology can be used for exploration of the dataset and business problem. This is often done by looking at the predictors and values that are chosen for each split of the tree. Often times these predictors provide usable insights or propose questions that need to be answered.

h) Step2: Data Preprocessing: Once the data has been explored then it is processed for other prediction algorithms. Because the algorithm is fairly robust with respect to a variety of predictor types (e.g. number, categorical etc.) and because it can be run relatively quickly decision trees can be used on the first pass of a data mining run to create a subset of possibly useful predictors that can then be fed into neural networks, nearest neighbor and normal statistical routines - which can take a considerable amount of time to run if there are large numbers of possible predictors to be used in the model.

i) Step3: Data Prediction: Although some forms of decision trees were initially developed as exploratory tools to refine and preprocess data for more standard statistical techniques like logistic regression. They have also been used and more increasingly often being used for prediction. This is interesting because many statisticians will still use decision trees for exploratory analysis effectively building a predictive model as a by product but then ignore the predictive model in favor of techniques that they are most comfortable with. Sometimes veteran analysts will do this even excluding the predictive model when it is superior to that produced by other techniques. With a host of new products and skilled users now appearing this tendency to use decision trees only for exploration now seems to be changing.

V. PERFORMANCES OF DECISION TREES

Decision trees are powerful and popular tools for classification and prediction. The attractiveness of decision trees is due to the fact that, in contrast to neural networks, decision trees represent *rules*. Rules can readily be expressed so that humans can understand them or even directly used in a database access language like SQL so that records falling into a particular category may be retrieved. In some applications, the accuracy of a classification or prediction is the only thing that matters. In such situations we do not necessarily care how or why the model works. In other situations, the ability to explain the reason for a decision is crucial. In marketing one has describe the customer segments to marketing professionals, so that they can utilize this knowledge in launching a successful marketing campaign. This domain expert must recognize and approve this discovered knowledge, and for this we need good descriptions. There are a variety of algorithms for building decision trees that share the desirable quality of interpretability. A decision tree can be used to classify an example by starting at the root of the tree and moving through it until a leaf node, which provides the classification of the instance.

Decision tree induction is a typical inductive approach to learn knowledge on classification. The key requirements to do mining with decision trees are:

*j) Attribute-value description***:** object or case must be expressible in item of a fixed collection of properties or attributes. This means that we need to discrete continuous attributes, or this must have been provided in the algorithm.

k) Predefined classes (target attribute values): The categories to which examples are to be assigned must have been established beforehand (supervised data).

l) Discrete Classes: A case does or does not belong to a particular class, and there must be more cases than classes.

m) Sufficient data: Usually hundreds or even thousands of training cases.

VI. Strengths and weakness of Decision Tree Technique

The strengths of decision tree methods are:

- Decision trees are able to generate understandable rules.
- Decision trees are able to handle both continuous and categorical variables.
- Decision trees perform classification without requiring much computation.

The weaknesses of decision tree methods:

- Decision trees are less appropriate for estimation tasks where the goal is to predict the value of a continuous attribute.
- Decision tree can be computationally expensive train. The process of growing a decision tree is computationally expensive. At each node, each candidate splitting field must be sorted before its best split can be found. In some algorithms, combinations of fields are used and a search must be made for optimal combining weights. Pruning algorithms can also be expensive since many candidate sub-trees must be formed and compared.
- Decision trees are prone to errors in classification problems with many class and relatively small number of training examples.

VII. Conclusion

Decision tree learning is one of the most important techniques in machine learning and data mining. It is a supervised technique that is often used when a disjunction of hypothese is required or when dealing (not exclusively) with categorical attributes. We build decision trees in order to capture underlying relaltionships in a datset. This can help us in classification and prediction as well as in data visualisation. It is preferrable largely because of the intuitive tree representationsof data that it produces. Many possible trees can be built that perfectly classify a given dataset. It is often preferrable to have small trees as they are easier to understand. Various algorithms exist to construct Decision Trees. All of them need some sort of criteria for selecting attributes to split data on. We looked at Information Gain, Gain Ratio and the use of the chi-squared distribution.

References

[1] CRS Report RL31798, Data Mining: An Overview, by Jeffrey W. Seifert.

[2] David B. Skillicorn, Parallel Data Mining.

[3] Chao-Tung Yang Shu-Tzu Tsai, Kuan- Ching Li, “Decision Tree Construction for Data Mining on Grid Computing Environments”, Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA’05).

[4] M. Berry and G. Linoff, *Data Mining Techniques,* John Wiley, 1997.

[5] *Technology Report*, Two Crows Corporation: “Introduction to Data Mining and Knowledge Discovery” Third Edition

[6] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules. Proc. of Conf. Very Large Data Bases (VLDB’94), pp. 487-499, Santiago, Chile Sept., 1994.

[7] Brin Sergey, Motwani Rajeev, Ullman Jeffrey, D. and Tsur Shalom. Dynamic item set counting and implication rules for market basket data. Proc. of ACM SIGMOD international conference on management of data, vol. 26, issue 2, pp 255-264, 1997.

[8] Bastide Yves. Taouil Rafik. Pasquier Nicolas. Stumme Gerd and Lakhal Lotfi. Mining Frequent Patterns with Counting Inference. In Proc. of ACM SIGKDD, pp 68-75, 2000.

[9] Sun Ken and Bai Fengshan. Mining Weighted Association Rules without Pre assigned Weights. In Proc. Of IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No. 4, pp 489-495, 2008.

[10] Jiawei Han and Michline Kamber . Data Mining Concepts and Techniques. Second edition, The Morgan Kaufmann series in Data Management Syatems, 2006.

[11] Cheung David W. Ngincent T. and Tam Benjamin W. Maintenance or discovered of knowledge: A case in multi-level association rules. In proceeding of Hong Kong research grant council, 1996.

[12] Show-Jane and Chen Arbee L.P. A Graph-Based Approach for Discovering Various Types of Association Rules. In Proc. of IEEE Transactions on Knowledge and Data Engineering. Vol. 13 No. 5, pp 839-845, 2001.

[13] Rajkumar N., Karthik M.R. and Sivanandam S.N. Fast Algorithm for Mining Multilevel Association Rules. In Proc. of IEEE Transactions on Knowledge and Data Engineering, vol. 02, 2003.

[14] Knowledge Discovery. KDD 1999: 425-429. 8888