Decision Tree approach to predict the Gender wise response on the volume of applications in government organizations for recruitment process

1Sunil Kumar Chhillar, 2Baljit Singh Khehra

1. Centre for Development of Advanced Computing, Mohali City, Punjab India

2. Baba Banda Singh Bahadur Engineering College, Fatehgarh Sahib City, Punjab India

Abstract

In this paper, we propose a practical method to predict the gender wise response on the volume of applications in government organization for recruitment process using Microsoft Decision Trees Algorithm which is a classification and regression algorithm provided by Microsoft SQL Server Analysis Services (SSAS) for use in predictive modeling of both discrete and continuous attributes.

The Microsoft Decision Trees algorithm approach is used builds a data-mining model by creating a series of splits, also called nodes, in the tree. The algorithm adds a node to the model every time an input column is found to be significantly correlated with the predictable column. The sample calculation shows that the proposed method is highly reliable that provides practical approach and tools for the wide use of data mining to predict results based on input, or attempt to find relationships among data or cluster data in previously unrecognized yet similar groups.

Earlier the predictions are available, better will be their utility for government i.e. provides valuable information with regard to procedure fixation, improve service delivery, reduce costs, redefine administrative processes etc.

Keywords- SSAS; Lift Chart, Decision tree

1. Introduction

Due to the rapid growth in the use of online data collection & processing and technologies in recent years, more and more data is collected and stored in a database. To gain useful information from these databases has become important issues faced by a number of organizations. The large potentials in the existing data banks need to be explored in order to transform these data/observable into valuable engineering information and knowledge. The key to these potentials can be found in data mining as a new emergent filed for prediction of gender wise response.

2. Decision Tree approach

The Microsoft Decision Trees algorithm is a hybrid algorithm that incorporates different methods for creating a tree, and supports multiple analytic tasks, including regression, classification, and association. It applies the Bayesian approach to learning causal interaction models by obtaining approximate posterior distributions for the models.

The methodology for assessing the information value of the priors needed for learning is based on the assumption of likelihood equivalence. Each case is assumed to have a single Bayesian prior network and a single measure of confidence for that network.

Using these prior networks, the algorithm then computes the relative posterior probabilities of network structures given the current training data, and identifies the network structures that have the highest posterior probabilities.

When the Microsoft Decision Trees algorithm creates the set of possible input values, it performs feature selection to identify the attributes and values that provide the most information, and removes from consideration the values that are very rare. The algorithm also groups values into bins, to create groupings of values that can be processed as a unit to optimize performance.

A tree is built by determining the correlations between an input and the targeted outcome. After all the attributes have been correlated, the algorithm identifies the single attribute that most cleanly separates the outcomes. This point of the best
separation is measured by using an equation that calculates information gain. The attribute that has the best score for information gain is used to divide the cases into subsets, which are then recursively analyzed by the same process, until the tree cannot be split any more.

The exact equation used to evaluate information gain depends on the parameters set when you created the algorithm, the data type of the predictable column, and the data type of the input.[7]

3. SQL Server Analysis Services

Microsoft SQL Server Analysis Services (SSAS) uses both server and client components to supply online analytical processing (OLAP) and data mining functionality for business intelligence solutions to achieve a top-down design method, which emphasize that the application decides data, and application have its own corresponding data. Using SSAS establish and manage cube, and analysis of it, allow developers define the data model called unified dimensional model (UDM) in one or more physical data sources to ensure the OLAP service procedure meet this design methods.

Fig.1. Physical Architecture (Analysis Services - Data Mining) [7]

4. Using Decision tree data mining technique

4.1 Decision Tress Algorithm Principle

The basic principle of the decision tree is recursively divide the data into subsets so that each sub-collection contains a similar state of target variables, whose properties are predictable. During each split of the tree, it is a must to evaluate the impact of predictable properties of input properties. When this recursive process is finished, the tree is also created.

4.2 Analysis and probability prediction

To predict the response, the data used is the simulation results of actual Online data colleted from the various government recruitments processes held during the last 3 to 4 years in the state of Punjab, includes data such as the department type (Education, Boards/Corporations etc.), post type (contractual or regular), recruitment year, and gender etc. SQL Server decision tree algorithm is used to implement data mining.

The tree (as shown in Figure 2) is created on the basis of department type and Post type factors that have different impacts on the response (As shown in Figure 3). From this response summary it is clear that the female response is more in government education sector against regular as well as contractual posts where as the male response is higher in government Board and Corporation & health sectors against regular posts. Figure 3 and 4 show that there is friendship between the access to Post type, Department type and gender. As shown in Figure 3, when the post type is contractual, the Proportion of gender='female' is higher than that of gender='male'; when the post type is Regular, the proportion of gender='male' is higher than the case of gender='female'.

Fig.2. Mining results of decision tree
4.3 Results

From the point of the whole process of data mining, the correctness of model verification plays a very important role. The lift chart (as shown in Figure 5) represents the analysis of the correctness of model for overall response. In the chart the results of mining model is displayed by a line, and simultaneously the chart also shows a number of other lines, these are the results produced by an ideal mode, with never wrong forecast. Figure 5 indicates the relationship between the two lines of different colors, and also give the satisfaction of the mining results is 64.05%, as shown in Figure 6.

The lift chart (as shown in Figure 7) indicates the relationship between the three lines of different colors along with the satisfaction of the mining results prediction for male gender as shown in Figure 8.

The lift chart (as shown in Figure 9) indicates the relationship between the three lines of different colors along with the satisfaction of the mining results prediction for female gender as shown in Figure 10.

**Fig.5. Lift chart (complete data)**

**Fig.6. Corresponding mining legend of lift chart**

**Fig.7. Lift chart (Male)**

**Fig.8. Corresponding mining legend of lift chart (Male)**

**Fig.9. Lift chart (Female)**

**Fig.10. Corresponding mining legend of lift chart (Female)**
5. Conclusion

This paper put forward the Model by Using Decision tree data mining technique; the predicted probability is obtained and verified the results using lift charts by carefully analyzing the correctness of model on the dataset on various recruitments in the government organizations of Punjab state, India. The results shows that the proposed method is highly reliable and improves the precision with a prior and likelihoods and predicted the gender wise response on the huge volume of applications in government organization for recruitment process.

So the mind of portfolio diagnosis and the methods in the paper is an effective approach that greatly simplifies the complexity of data mining operations, so that ordinary programmers are able to complete data mining operations, thus it provides practical approach and tools for the wide use of data mining and provides valuable information on the basis of this analysis to the government with regard to procedure fixation, improve service delivery, reduce costs, redefine administrative processes.

6. References


[8] Zhaohui Tang, Jamie Maclellan, Peter Pyungchul Kim, Building Data Mining Solutions with OLE DB for DM and XML for Analysis, SIGMOD Record, Vol. 34, No. 2, June 2005