The Analysis of English Sentence Components Based on Decision Tree Classification Algorithm

Hui Zhang*
Department of Primary Education, Shangrao Preschool Education College, Shangrao, China
*Corresponding author: wztgeducation@sina.com

Abstract. Decision tree is an important classification method in data excavation technology. It is a predictive analysis model expressed in the form of a tree structure (including binary trees and poly trees). The decision tree method is a more general classification function approximation method. It is an algorithm commonly used in predictive models to find some potentially valuable information by purposefully classifying a large amount of data. In this article, the author tries to analyze the English sentence components based on the decision tree classification algorithm. The author starts with the decision tree, extracts the decision tree rules, and generates a classifier by effectively sorting the decision tree rules, and applies it to classification prediction.

Keywords: Decision Tree; Classification Algorithm.

1. Introduction of Theory of Data Excavation

The so-called data excavation (also known as knowledge discovery in databases) refers to the extraction of implicit and unknown data from a large amount of incomplete, noisy, fuzzy, random mass data, or large databases or data warehouses. It contains non-trivial information or patterns with potential application value. The process of data excavation is referred to determining the purpose of mining, data preparation, data mining, pattern evaluation and knowledge representation as shown in Fig.1.

![Figure 1. The Specific Process of Data Excavation](image)

And the so-called classification is to establish a classifier for the description rules of the known existing categories, and then judge and classify the observation values of the unknown new cases (Figure 2). Data classification is one of the main content of data mining, mainly through the analysis of training data samples, to produce an accurate description of the category. This category is usually composed of classification rules, which can be used to classify and predict future data. The commonly used classification methods include: decision tree classification, association classification, neural network, Bayesian classification method, etc (Figure 3). Typical algorithms based on decision tree classification are: ID3 algorithm, C4.5 algorithm, PART algorithm, CABRR algorithm and so on. Each classification method needs to be evaluated with certain indicators. The comparison and evaluation criteria of commonly used classification algorithms are as follows: a) Prediction accuracy. It refers to the model's ability to correctly predict the class label of new or previously unseen data. b) Comprehensibility. It refers to the explanatory power of the model. c) Scalability speed. It refers to the computational cost of generating and using the model, as well as the speed of building the model. d) Robustness. It refers to the ability of the model to correctly predict given noisy data or data with missing values. There are two steps of data classification.

Step one: Build a mode to describe a given data set or concept set (training set for short).

Step two: Use the mode to classify the data. Including evaluating the classification accuracy of the mode and classifying tuples with unknown class labels according to the mode.
2. Decision Tree

Decision tree is an important classification method in data mining technology. It is a predictive analysis model expressed in the form of a tree structure (including binary trees and poly trees). The instance is classified by arranging the instance from the root node to a leaf node; the leaf node is the category to which the instance belongs; each node on the tree illustrates the test of a certain attribute of the instance, and each subsequent branch of the node corresponds to the possible value for the attribute.

The decision tree method is a more general classification function approximation method. It is an algorithm commonly used in predictive models to find some potentially valuable information by purposefully classifying a large amount of data. The origin of decision tree is CLS (Concept Learning System). CLS was developed by Hunt, Marin and Stone to study human conceptual models. It was proposed in 1966. This model has laid a good foundation for the development of many decision tree algorithms. In 1984, L. Breiman et al. proposed the CART (Classification and Regression Tree) algorithm. In 1986, J.R. Quinlan proposed the ID3 algorithm. In 1993, J.R. Quinlan proposed the C4.5 algorithm, which overcomes some of the shortcomings of the ID3 algorithm. In 1996, M. Mehta and R. Agrawal and others proposed a high-speed scalable supervised learning classification method SLIQ (Supervised Learning In Quest). In the same year, J. Shafer and R. Agrawal and others proposed a scalable parallel inductive decision tree classification method SPRINT (Scalable PaRallelizable Induction of Decision Trees). In 1998, R. Rastogi et al. proposed a classification algorithm PUBLIC (A Decision Tree that Integrates Building and Pruning) that combines tree building and pruning.

It consists of two elements as shown in Figure 4: nodes and branches. In the finally generated decision tree, each internal node represents an attribute of the data set, each branch represents a test output of the attribute, each leaf node represents the divided category, and the top node is the root node.

The decision tree generation process: It is mainly divided into two steps: one is the spanning tree, and the other is the tree pruning.

Tree pruning: the pruning of trees. The most commonly used pruning techniques include pre-pruning and post-pruning.
3. Case study on English Sentence Components Analysis by Decision Tree

The decision tree is established through recursive partitioning, which is an iterative process of dividing data into different small parts. If the following situation occurs, the decision tree will stop splitting: each data of the group of data has been classified into the same category; There is no way for this group of data to find new attributes for node segmentation; There is no unprocessed data in this group of data. Decision tree learning may encounter the problem of model over-fitting. Over-fitting refers to the over-training of the model, which causes the model to remember not the generality of the training set, but the local characteristics of the training set. How to deal with over fitting? Prune the decision tree. There are several solutions for tree pruning, mainly pruning first and pruning later.

The decision tree is applied in the sentence components analysis of English Grammar. It helps language learners understand and figure out the structure of each sentence. The author presented some examples to show the application of decision tree.

a) The apple might hit the man.

b) Mary is fond of literature but tired of statistics.

c) The detective went out and the mysterious man came in.

d) The dog that he keeps bites.

Figure 5. The Decision Tree of English Sentences
4. Conclusions

The decision tree could clearly display the decision-making process of the whole decision-making problem in solving problems. When applied to complex multi-stage decision-making, the stages are obvious and the levels are clear. There will be many problems in the process of decision tree development. We need to figure them out and solve them! In this way, the method will help language learners make the components structure clearly.

References