Improving the Classification accuracy of Noisy Dataset by Effective Data Preprocessing

Abstract

Decision tree is a technique commonly used in data mining. Issues in decision tree algorithms are working with continuous attributes and missing values, avoiding over fitting, super attributes. Handling noisy data is the challenging factor in data mining research. Noisy data is meaningless data. It unnecessarily increases the amount of storage space required and can also adversely affect the results of any data mining analysis. Predicting the result from such noisy data is the complicated factor. The commonly used algorithm for classification problems are decision stump, ensemble models, SVM, and decision tree algorithms. The performance of the algorithm resulted in lower accuracy when comparing with the noiseless data result. Thus in this paper, data is collected and noise is added to the data, and then it is preprocessed for handling missing values. The preprocessed data is then provided as the input for the feature selection technique. Most relevant features are selected using correlation based subset feature selection technique. The selected features are provided as the input of Credal C4.5 algorithm and decision tree is constructed. The result is analyzed with various data with (5,10,20,30)% noise level. This technique improves the performance of the algorithm with (1-5)% improvement in
accuracy compared to the existing result.

References


Index Terms

Computer Science

Information Sciences
Keywords

Classification, Noisy Data, Feature Selection, Data Preprocessing.