Abstract

Data mining plays an important role in processing large volumes of data. It refers to the process of obtaining knowledge from raw data. Classification is the most widely used data mining techniques, which employs some set of pre-classified samples to develop a model called a classifier. Many researches showed that C4.5 algorithm need to be improvised to maximize accuracy, handle large amounts of data, where C5.0 is the improved version. The major goal of the classification technique is to predict the target class accurately for each case in the data. The main objective of this research work is to predict diseases using classification algorithms such as Decision trees, C5.0 and Bayesian Networks. The performance of classification algorithms is compared using the datasets, Breast cancer and Heart disease. The experimental results are compared based on different performance parameters like dataset scalability, accuracy and error rate values. The research shows that in terms of scalability Bayesian networks algorithm was proved to have more accuracy rate and less error rate than the C5.0 algorithm.
References

1. Soumen Chakrabarti, Earl Cox, Eibe Frank, Ralf Hartmut Güting, Jaiwei Han, Xia Jiang, Micheline Kamber, Sam S. Lightstone, Thomas P. Nadeau Richard E. Neapolitan, Dorian Pyle, Mamdouh Refaat, Markus Schneider, Toby J. Teorey, Ian H. Witten, “Data Mining-Know it all”, Morgan Kaufmann Publishers, 2009


7. FUZZY SETS AND SYSTEMS, Elsevier An International Journal in Information Science and Engineering


16. Learning Bayesian Network Model Structure from Data Dimitris Margaritis May 2003 CMU-CS-03-153 School of Computer Science Carnegie Mellon University Pittsburgh, PA 15213


Index Terms

Computer Science

Information Sciences

Keywords

Classification, C5.0, Bayesian Networks, Decision tree, Disease rules, Disease Prediction