Abstract

Medical Data mining is the process of extracting hidden patterns from medical data. This paper presents the development of a hybrid model for classifying Pima Indian diabetic database (PIDD). The model consists of two stages. In the first stage, the K-means clustering is used to identify and eliminate incorrectly classified instances. The continuous data is converted to categorical form by approximate width of the desired intervals, based on the opinion of medical expert. In the second stage a fine tuned classification is done using Decision tree C4.5 by taking the correctly clustered instance of first stage. Experimental results signify the cascaded K-means clustering and Decision tree C4.5 has enhanced classification accuracy of C4.5. Further rules generated using cascaded C4.5 tree with categorical data are less in numbers and easy to interpret compared to rules generated with C4.5 alone with continuous data. The proposed cascaded model with categorical data obtained the classification accuracy of 93.33% when compared to accuracy of 73.62% using C4.5 alone for PIMA Indian diabetic dataset.
References

- Editorial, Diagnosis and Classification of Diabetes Mellitus, American Diabetes Association, Diabetes Care, vol 27, Supplement 1, (Jan 2004).
- B. M Patil, R. C Joshi, Durga Tosniwal, Hybrid Prediction model for Type-2 Diabetic Patients, Expert System with Applications, 37, 8102-8108 (2010).
- Asha Gowda Karegowda and M. A. Jayaram, Cascading GA & CFS for Feature Subset Selection in Medical Data Mining, International Conference on IEEE International Advance Computing Conference (IACC’s09), Thapar University, Patiala, Punjab India (Mar 2009).

Index Terms

Computer Science
Artificial Intelligence
Keywords
K-means Clustering  Categorical Data  Rule Based Classification  Decision Tree C4. 5
Pima Indian Diabetics