Abstract

This paper develops the cascaded models for classification of PIMA Indian diabetes database. The k-nearest neighbour method is used to impute the missing data and the processed data is used for further classification. This is done in two steps, in first step k-means clustering algorithm is used for extracting hidden patterns in data set then in second step the classification is done by using suitable classifier. k-means algorithm combined with artificial neural network classifier and k-means algorithm combined with logistic regression classifier achieve classification accuracy above 98%.

References

2. A. G. Karegowda, M. A. Jayaram, Integrating Decision Tree and ANN for Categorization of Diabetics Data, International Conference on Computer Aided Engineering, December 13–15,
IIT Madras, Chennai, India (2007).
7. Gustavo E. A. P. A. Batista and Maria Carolina Monard, University of Sao Paulo, A Study of k- Nearest Neighbour as an Imputation Method.

Index Terms

Computer Science  Information Sciences

Keywords

Missing data, Clustering, Classification