Objectives

Feature selection in data sets is an important task allowing to alleviate various machine learning and data mining issues. The main objectives of a feature selection method consist on building simpler and more understandable classifier models in order to improve the data mining and processing performances. Therefore, a comparative evaluation of the Chi-square method, recursive feature elimination method, and tree-based method (using Random Forest) used on the three common machine learning methods (K-Nearest Neighbor, naïve Bayesian classifier and decision tree classifier) are performed to select the most relevant primitives from a large set of attributes. Furthermore, determining the most suitable couple (i.e., feature selection method-machine learning method) that provides the best performance is performed.

Materials and methods

In this paper, an overview of the most common feature selection techniques is first provided: the Chi-Square method, the Recursive Feature Elimination method (RFE) and the tree-based method (using Random Forest). A comparative evaluation of the improvement (brought by such feature selection methods) to the three common machine learning methods (K- Nearest Neighbor, naïve Bayesian classifier and decision tree classifier) are performed. For evaluation purposes, the following measures: micro-F1, accuracy and root mean square error are used on the stroke disease data set.

Results

The obtained results show that the proposed approach (i.e., Tree Based Method using Random Forest, TBM-RF, decision tree classifier, DTC) provides accuracy higher than 85%, F1-score higher than 88%, thus, better than the KNN and NB using the Chi-Square, RFE and TBM-RF methods.

Conclusion

This study shows that the couple - Tree Based Method using Random Forest (TBM-RF) decision tree classifier successfully and efficiently contributes to find the most relevant features and to predict and classify patient suffering of stroke disease.