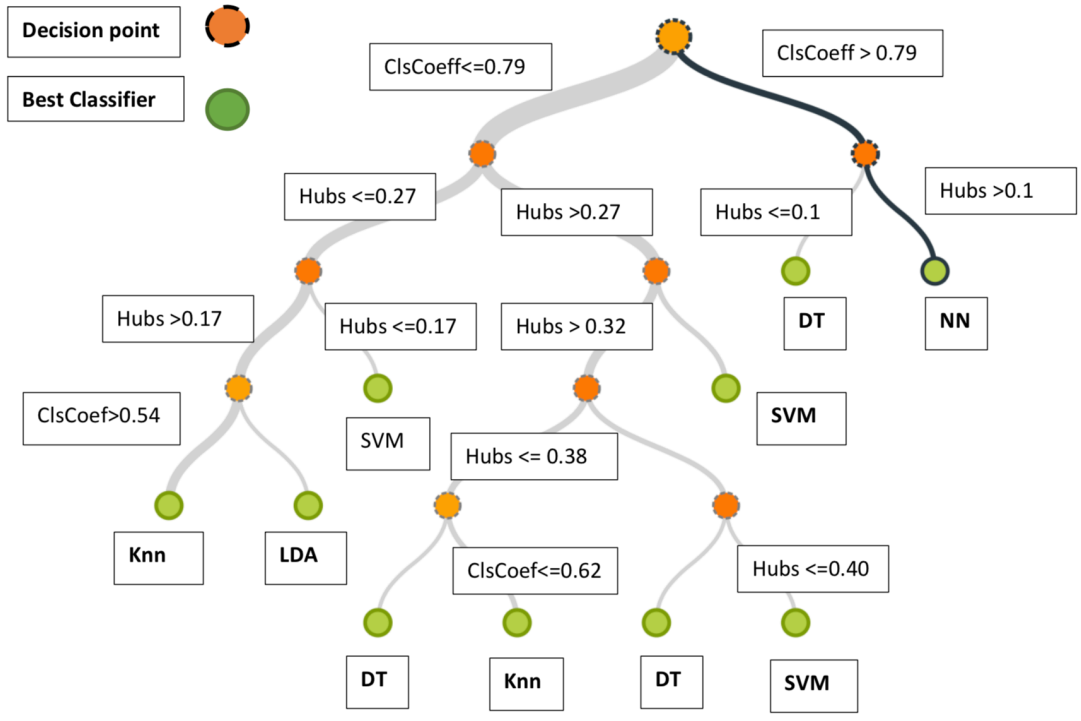
Importance of Network Metrics in Classification

Dieudonne Ouedraogo

Abstract

In machine learning, the performance of a classifier is intrinsically related to the task. The structure between data points within the dataset plays an important role. In this paper, we explore the usage of network metrics to describe the selection of a machine learning algorithm for a classification task concerning a specific dataset. A dataset is transformed into a graph representation based on the ϵNN algorithm. A data point is a node, and an edge exists between two points i,j if $d(i,j) < \epsilon$. A post-processing step is applied to the graph, pruning edges between examples of different classes. The structural information such as density, clustering coefficient, and hubs are extracted. Various data sets are collected, their network metrics are computed. A predictive model is built to investigate the possible relationship between networks characteristics and the classifier used to be used on the machine learning task. Results show that network metrics such as clustering coefficient, hubs,density are very informative in predicting the classifier to be used on the task. For example, clustering coefficient greater than 0.79 or hubs greater 0.1, neural network are the best performers; for hubs less than 0.1, using decision trees gives the highest accuracy on the predictions.



Decision Tree with network metrics as variables and algorithms as outputs