Abstract

This presents an investigation into the usefulness of rough set theory in the context of authorship attribution using writing style. The problem was setup as a standard supervised machine learning problem. The rough set based feature subset computation techniques reduced the dimensionality of the feature space from 346 conditional attributes to an average of 8 features. Experiments were performed experiment using five different subsets of the original attributes computed using rough sets techniques with the results showing that the rough set based techniques improved the performances of neural network (NN) and Support Vector Machines (SVM) models. The overall classification accuracy increased from 8.712 % for on the baseline data to 50.505 % for the NN and from 7.197 % to 28.662 % for the SVM model. The improvements in performance compared to the baseline model are evidenced across all other performance metrics used. However, the NN model performed generally better than the SVM model.


Index Terms

Computer Science  Information Sciences

Keywords

Stylometry, Feature Selection, Neural Networks, Support Vector Machines, Supervised Machine Learning.