Infomation based supervised and semi-supervised feature selection

Sang-Keun Lee, Seung-Joon Yi, Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University
sklee@bi.snu.ac.kr, sjlee@bi.snu.ac.kr, btzhang@cse.snu.ac.kr

Abstract

We merge the results from both of supervised and semi-supervised feature selection techniques. The method was applied to the five datasets from NIPS feature selection competition. As a preprocessing step, we firstly discretize each training dataset using EM algorithm. Then, we filter the discretized dataset based on the MI (mutual information) value of each feature with respect to the class variable, selecting the features having higher MI value than background noises. The background noise level corresponds to the average MI values of randomly-generated feature sets. The filtered five datasets have 2,882 (28.8%), 1,958 (39.2%), 5,032 (25.2%), 5,951 (6.0%), and 111 (22.2%) features, respectively (the same order as on the competition website).

In the supervised feature selection approach, we use two measures of relevance, the *significance* ($I(F_1; Class|F_2)$) and the *independence* ($I(F_1; F_2|Class)^{-1}$) of feature. By a bidirectional greedy search, the features maximizing both of these measures are chosen. In the semi-supervised approach, we deployed an agglomerative clustering technique using the independence between features ($I(F_1; F_2)^{-1}$) as a distance metric. From each cluster, the most relevant ($I(F; Class)$) features are selected.

By the supervised method, we extracted 185 (1.9%), 362 (7.2%), 827 (4.1%), 239 (0.2%), and 39 (7.8%) features from the five datasets, respectively. Using the semi-supervised approach, 289 (2.9%), 186 (3.7%), 504 (2.6%), 597 (0.6%), and 12 (2.4%) features were selected. For each dataset, we merged the above feature sets for classification. Consequently, 417 (4.2%), 467 (9.3%), 1,017 (5.0%), 766 (0.8%), and 40 (5%) features were selected for the five datasets. As a classification method, the support vector machine, naive Bayes classifiers, and Adaboost with naive Bayes were adopted. As a result, the classification performance using the merged feature sets surpassed the cases of using only one method.