ABSTRACT OF THE MASTER THESIS

SINA TABAKHI

OCTOBER 2013

Practical applications of data classification that produce huge volume datasets have been growing rapidly nowadays. Many of the datasets have large numbers of features and relatively small numbers of patterns, many of whose features are often irrelevant or redundant. It may have a negative impact on the performance of the classification. One important technique to overcome this problem is feature selection. The purpose of feature selection is to select a subset of useful features from the input feature set to improve classification accuracy. The exhaustive search for finding the optimal feature subset is computationally intractable in a reasonable time. Therefore, feature selection arises a major challenging problem for pattern recognition and machine learning.

Four new filter based feature selection methods are proposed in this thesis. The search space is represented as a weighted graph, in the first method, and the optimal feature subset is determined using ant colony optimization algorithm. Applying a swarm intelligence based technique in the filter approach, using no learning algorithms, is the main advantage of the proposed method. Moreover, based on the similarity between features, which leads to the minimization of the redundancy, the feature relevance is computed. The other methods have been proposed based on the weaknesses of the first method. In the second proposed method, the relevance of each feature is individually assessed and incorporated into the feature selection process. In the third proposed method, incremental feature selection is introduced wherein the similarity of the new feature is computed with a subset of previously selected features. In other words, a new heuristic information measure is presented that considers the similarity between subsets of features to enhance the redundancy reduction process in the proposed method. Finally, the concept of simulated annealing is applied to control the randomness of the ant colony optimization algorithm in the fourth method.

The performance of the proposed methods has been compared to 11 well-known univariate and multivariate feature selection methods using different classifiers. The experimental results show not only the efficiency and effectiveness of the proposed methods but the improvements over previous related ones. Furthermore, compared to the others, the proposed methods find a smaller number of features.