Classification Algorithm Selection: Performance Metric, Evaluation Method, and FAMCDM

Yi Peng

School of Management and Economics, University of Electronic Science and Technology of China, Chengdu, 610054, P. R. China

Abstract

As a major data mining task, many classification algorithms have been developed and applied to various applications. The performances of classifiers vary with different performance measures and under different circumstances. As the No Free Lunch (NFL) theorem states, “if algorithm A outperforms algorithm B on some cost functions, then loosely speaking there must exist exactly as many other functions where B outperforms A”. There exists no single classification algorithm that could achieve the best performance for all measures. Many previous studies assess the performance of classifiers using only a couple of traditional measurements (e.g., the classification accuracy and the area under the receiver operating characteristic curve (AUC)), which have certain limitations. How to provide a comprehensive assessment of classifiers and recommend an adequate classifier (or set of classifiers) is an important and understudied area.

The algorithm evaluation or algorithm selection problem is an active research area in many fields, such as artificial intelligence, operations research, and machine learning. Since the algorithm selection task needs to examine several criteria, it can be modeled as a multiple criteria decision making (MCDM) problem and MCDM methods can be used to systematically choose appropriate algorithm(s).

This talk summarizes three aspects of the author and her colleagues’ recent works: designing performance metric for classification algorithm evaluation; introducing MCDM methods to rank classifiers; and developing a fusion approach to reconcile conflicting rankings generated by different MCDM methods.

There are an extensive number of measures for classification. These measures have been introduced for different applications and to evaluate different things. Thus we designed a performance metric to evaluate the merit of classification algorithms using a broad selection of classification algorithms and performance measures. The basic idea of this performance metric is similar to ranking methods, which use experimental results generated by a set of algorithms on a set of datasets to rank those algorithms. It resembles the significant wins (SW) ranking method by conducting pairwise comparisons of classifiers using *t* tests with a significance level of 0.05. The metric was assessed using 13 well-known classification methods over 11 public-domain data sets from the NASA Metrics Data Program (MDP) repository.

Ranking of classification algorithms normally need to examine several criteria, such as accuracy, computational time, and misclassification rate. Therefore algorithm selection can be modeled as multiple criteria decision making (MCDM) problems. Some existing MCDM methods are able to rank classifiers based on multiple performance measures and take the preferences of users into the ranking process. We proposed to use MCDM methods (i.e., DEA, ELECTRE I, TOPSIS, and PROMETHEE II) to evaluate and rank a selection of classification algorithms using a set of performance measures. Since the preferences of the decision maker (DM) play an important role in algorithm evaluation and selection, we involved user's preferences during the ranking procedure by assigning weights to evaluation criteria.

While the rankings of algorithms provided by different MCDM methods are in agreement sometimes, there are situations where MCDM methods generate very different results. To resolve this disagreement and help decision-makers pick the most suitable classifier(s), we proposed a fusion approach to produce a weighted compatible MCDM ranking of multiclass classification algorithms. The key of the fusion approach is to determine a set of weights for different MCDM methods by finding a compromise solution that has the minimum overall difference between the optimal ranking score and all the available ranking scores. Experimental studies using four public-domain multiclass datasets from four different application domains were conducted to validate the proposed fusion approach. The rankings generated by the four MCDM methods for the four multiclass datasets are quite different at first. After applying the fusion approach, the secondary rankings of the MCDM methods are in strong agreement. Specifically, rankings of classifiers for two datasets are identical and only slightly different for the other two datasets. The experimental results indicate that the fusion approach proposed in this paper can provide a compatible ranking when different MCDM techniques yield conflicting results.

Keywords: algorithm selection, algorithm evaluation, performance metric, classification, multiple criteria decision making (MCDM)