The Optimal Diagnostic Decision Sequence

Chih-Lin Chi[1], W. Nick Street[2]

[1]Health Informatics Program, University of Iowa

[2]Management Sciences Department, University of Iowa

Abstract

We describe a data mining model for constructing an optimal diagnostic sequence that assists cost-effective sequential decisions. We use heuristic search, i.e., hill climbing and genetic algorithms (GAs), and the evaluation function of cost-based Mean Accuracy Gain (cMAG), which is provided by SVM classifiers, to find this optimal sequence. GA can find a good sequence because of the ability to escape from local optima.

Background

Triage sorts patients and provides appropriate resources according to patients' risks. It is applied to diagnose patients with tests sequentially. An appropriate sequence of tests can lead to quick diagnoses, high diagnostic performance, and cost savings. The disease-specific diagnostic sequences are usually constructed by meetings of several domain experts.

This study provides an alternative way to construct this sequence using feature selection. Tests selection and their order in a sequence are formulated as an optimization problem. Similar to wrapper-based feature selection, we use a group of classifiers to find cost-effective features (tests) but take the order of time of introducing new features into account. Classifiers are used sequentially based on this order to provide an evaluation function. The number of possible sequences is very huge. We use heuristic search to solve this optimization problem based on the evaluation function.

Methods

We use a classifier to evaluate the performance of a test point and its prior tests along the sequence, using accuracy as the performance index. The basic idea is to find the sequence with good performance for every test point. The evaluation function is the mean of the ratio of accuracy gains and test costs

$$cMAG = \underset{i}{mean}\left(\frac{acc_i - acc_{base}}{c_i}\right)$$

where acc_i is the accuracy of the classifier for test point i, acc_{base} is the accuracy of the classifier without any tests (base information, e.g., symptoms), and c_i is the cost of test point i.

We use a neighborhood function (the combination of 2-exchange neighbor and random key) to create permutations based on the local changes in tests selection and their order. Heuristic search then finds the best permutation based on *cMAG*.

Results

We apply this model to heart disease data from the UCI Machine learning repository. There are nine possible tests in this heart disease diagnostic dataset. Evaluation is performed using five 10-fold cross-validation runs. The base accuracy is 0.7769, and the accuracy of the classifier using all tests is 0.8282. The benchmark, random search, finds a sequence consisting of 6 tests. *cMAG* is 1.13 and accuracies of tests are 0.80, 0.77, 0.79, 0.79, 0.80, and 0.79. Hill climbing finds a sequence with 5 tests, *cMAG* of 1.7478 and accuracies of 0.78, 0.82, 0.83, 0.8, and 0.83. Genetic algorithm (GA) finds the best sequence with 7 tests. *cMAG* is 2.2183 and accuracies are 0.78, 0.83, 0.83, 0.84, 0.83, 0.83, and 0.84. Accuracies of most points on the sequence found by GA are better than the accuracy of the classifier using all 9 tests.

To use the optimal sequence, we need to build a series of predictive models. Each predictive model is trained with base information and certain tests based on the sequence. Each predictive model can assist diagnostic decision after receiving a test result along the optimal sequence.

Conclusion

This study uses classifiers and heuristic search to find the best sequence with good diagnostic accuracy for every test point. There are two meanings for this optimal sequence. First, fewer tests can be used to diagnose more accurately, using less time and fewer resources. Thus, triage decision can be facilitated. Second, this project extends the use of feature selection to medical diagnostic problems.