A data mining technique for risk-stratification diagnosis

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 sequential diagnosis, called the Optimal Decision Path Finder (ODPF), which is built based on the idea of risk stratification. A filter was used to stratify patients depending on ease of diagnosis, and a series of patient-specific classifiers was built to diagnose with confidence while reducing exam cost. Results show that applying stratification to data mining can improve the diagnostic performance and reduce waste of medical resource. This resulting model can assist the physician in triage decisions.

Background

Risk stratification, which means the sorting of patients based on the severity of illness, is an important issue in medical diagnosis because it can help to reduce usage of beds, equipment, and other medical resources. While clinical guidelines provide a sequence of diagnostic exams, physicians still need to decide whether a patient needs further exams based on experience. In some cases, the symptoms and evidence from a subset of available tests are sufficient to support particular diagnosis with high confidence. In other words, these patients are easy-diagnosis cases. For hard-diagnosis cases, patients may need to receive all possible tests, and the diagnosis decision may still be difficult. Our solution to this problem is ODPF, a stratification filter that applies to an individualized sequence of tests ordered by cost-effectiveness. It speeds up accurate diagnosis and improves the performance of the prediction model.

Methods

ODPF is based on the idea on lazy-learning prediction models. An instance weight based on the similarity to the query patient is applied to each case during learning, resulting in a sequence of patient-specific classifiers. The sequence of tests is decided by a greedy-search feature selection method using the evaluation function of performance gain per dollar (cost-effectiveness). This function promotes the selection of either very simple and inexpensive tests or high performance-improvement tests with a reasonable price. This evaluation function accelerates cost-effective diagnosis. As a result, many unnecessary tests can be avoided. The greedy search is performed at each step in the diagnosis process. Finally, a complete medical testing sequence can be determined. A confidence threshold, such as the 90% predictive probability, was used as the filter to stratify the ease of diagnosis of patients. During the iterative process, the filter can determine when there is enough evidence to diagnose accurately.

Results

We apply this model to a heart disease data from the UCI Machine learning repository. There are nine possible exams in this heart disease diagnosis dataset. Evaluation is performed using five 10-fold cross-validation runs. The proportions of patients who receive no tests to all tests (ease-of-diagnosis strata) are 0.136, 0.047, 0.059, 0.081, 0.077, 0.061, 0.058, 0.049, 0.041, and 0.39. The first stratum contains the easiest-diagnosis patients, and the last stratum contains the hardest-diagnosis patients. Around 1/3 patients are very hard to diagnose. The corresponding accuracies are 0.913, 0.905, 0.911, 0.917, 0.894, 0.904, 0.924, 0.848, 0.839, and 0.756. The total accuracy from all cases is 0.844, which is significant better than the accuracy of patients who receive all tests (0.83). As expected, there is a trend that accuracy will decrease when approaching the “hard end” of the ease-of-diagnosis strata. The average number of required exams of all patients is 5.57, and the cost saving is about 46%.

Conclusion

The ODPF can help three types of decision-making. First, it determines the optimal testing sequence based on cost and performance gain. Second, a stratification filter can help to decide whether or not there is enough evidence for accurate diagnosis. Third, ODPF provides diagnosis assistance for each stratum. This triage decision support model can help to improve the accuracy of diagnosis and reduce the waste of medical resources.