Chapter 9
Machine Learning for Clinical Data Processing

Guo-Zheng Li
Tongji University, China

ABSTRACT

This chapter introduces great challenges and the novel machine learning techniques employed in clinical data processing. It argues that the novel machine learning techniques including support vector machines, ensemble learning, feature selection, feature reuse by using multi-task learning, and multi-label learning provide potentially more substantive solutions for decision support and clinical data analysis. The authors demonstrate the generalization performance of the novel machine learning techniques on real world data sets including one data set of brain glioma, one data set of coronary heart disease in Chinese Medicine and some tumor data sets of microarray. More and more machine learning techniques will be developed to improve analysis precision of clinical data sets.

INTRODUCTION

Clinical data has increased dramatically with the rapid development of computing technology, which produces great need from intelligent techniques, especially novel machine learning techniques. During clinical data processing, there are great challenges to face for machine learning researchers:

- There are complex structures in the clinical data sets, e.g. for a patient, if he feel Stomachache, we need to record when, where and how he felt. In current data sets, we often ignore some of the elements like the exact occasion, which hurt the decision support from intelligent techniques, since different moments may imply different causes.
- The data set is in sparse representation, when there are many symptoms to describe

DOI: 10.4018/978-1-60960-483-7.ch009

one disease. We use some symptoms to describe some patients, while others to describe others. Sparse representation makes the features be high dimensional, while the samples be few, which hurts the modeling of machine learning techniques.

- The data sets are small sample due to the cost. Though support vector machines are developed to overcome this problem, but there is still a gap between the machine learning techniques with the human experience. Human doctors infer experience from only one or a few cases, while computer cannot. Computers make errors when there are only few cases due to the the law of large numbers.
- There are imbalanced problems between the positive class and the negative one. We often focus on positives, but they are often less than negatives. Computers often biases to the bigger class, i.e. the negative. This hurts the generalization performance and reduces the forecast accuracy of positive cases.
- There are noise and bias in the existing cases. We know different clinicians may produce different judgments on one patient, so what we collected may not objective, which also hurts the modeling of these applications.
- There are many classes or labels for one case, i.e. one patient may have two more diseases. This is also a challenge for machine learning techniques; the existing techniques are excellent on binary classification. Novel techniques are needed to develop to solve different problems.
- Understanding needs to be improved in the existing techniques, we not only need high forecast accuracy, we also want to know why the techniques produce such results, we need to interpret the results and the intrinsic principles behind the problem. Unfortunately, most of the popular machine learning techniques is black box; this may need further works for machine learning researchers after they developed good techniques with high performance.

To face the above challenges in clinical data processing, this chapter describes the novel techniques of support vector machines, ensemble learning, feature selection, multi-label learning and feature reuse by using multi-task learning. Applications include degree forecast of brain glioma, syndrome classification of inquiry diagnosis for coronary heart disease in Chinese medicine and tumor classification of microarray data sets.

Support vector machines (SVMs) are excellent in solving small sample problems (Boser, Guyon, & Vapnik, 1992; Vapnik, 1995). Since 1990s, they have developed rapidly and reached state-of-the-art performance in small sample problems (Cristianini & Shawe-Taylor, 2000; Chen, Lu, Yang, & Li, 2004). Compared to artificial neural networks (ANNs), SVMs have better generalization performance and can obtain a global optimal solution. At the same time, a type of feature selection algorithms based on SVMs named embedded algorithms has been proposed (Li, Yang, Liu & Xue, 2004; Lal, Chapelle, Weston & Elisseeff, 2006; Li, Meng, Yang & Yang, 2009) to solve the problems involving many irrelevant and redundant features such as symptom selection and tumor categorization. These embedded algorithms are designed to select features efficiently, but, compared to wrapper methods (Kohavi & George, 1997) the accuracy is sacrificed in some degree. Furthermore, subset generation methods used in most embedded algorithms are methods like sequential backward feature selection which cannot effectively handle combined features in the feature selection procedure.

In the literature, floating search (Pudil, Novovicova & Kittler, 1994) which is a heuristic feature subset generation algorithm, has been proved one of the best subset generation algorithms (Jain & Zongker, 1997; Kudo & Sklansky, 2000)