Chapter 4

Data Mining Based on Intelligent Systems for Decision Support Systems in Healthcare

Loris Nanni[1], Sheryl Brahnam[2], Alessandra Lumini[1], and Tonya Barrier[2]

[1] DEIS, IEIIT—CNR, Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
loris.nanni@unibo.it, alessandra.lumini@unibo.it
[2] Computer Information Systems, Missouri State University,
901 S. National, Springfield, MO 65804, USA
sbrahnam@missouristate.edu
tonyabarrier@missouristate.edu

Abstract. In this paper we make an extensive study of Artificial Intelligence (AI) techniques that can be used in decision support systems in healthcare. In particular, we propose variants of ensemble methods (i.e., Rotation Forest and Input Decimated Ensembles) that are based on perturbing features, and we make a wide comparison among the ensemble approaches. We illustrate the power of these techniques by applying our approaches to different healthcare problems. Included in this chapter is extensive background material on the single classifier systems, ensemble methods, and feature transforms used in the experimental section.

Keywords: rotation forest, input decimated ensembles, multiclassifier systems, decision trees, medical decision support systems.

1 Introduction

Across the globe, healthcare is suffering from financial pressures. When making healthcare decisions, administrators and physicians must analyze both clinical and financial information. Many organizations are asking how they can address the problem of maintaining quality healthcare while simultaneously lowering costs to remain competitive. To address this problem, healthcare organizations are starting to make extensive use of state-of-the-art data mining technologies.

Data mining uses advanced statistical methods, database technology, artificial intelligence, pattern recognition, machine learning, and data visualization to discover hidden patterns in data. Two potential uses of data mining in medicine that could reduce costs significantly include predicting patient reactions to drugs and identifying patients who might best benefit from adjuvant chemotherapy (by comparing, for example, patients who continue to be disease free after five years with patients who develop metastases within five years). By properly managing patients today, tomorrow's problems and costs can be reduced. It may soon be

S. Brahnam & L.C. Jain (Eds.): Adv. Comput. Intell. Paradigms in Healthcare 5, SCI 326, pp. 45–65.
springerlink.com

possible, given the rise of telemedicine, for data mining methods to be stored on centralized servers accessible to physicians around the world. By submitting relevant information about patients and outcomes, servers would be able to offer expert recommendations that could continuously be refined. In these and other ways, data mining has the potential of helping healthcare organizations achieve their goal of maximizing quality care while minimizing costs.

With scientists continuously producing more information than can be processed, a phenomenon frequently referred to as data avalanche, the potential offered by data mining will only continue to grow. What needs to be developed today to begin realizing this potential are new classification methods that are highly flexible and that offer reasonable accuracy in prediction. In other words, we need general purpose classification methods that are capable of handling a wide variety of medical problems, that compare well with human expertise, and that compete with less flexible state-of-the-art methods that have been crafted for very specific problems. In addition, these new classification methods need to be able to integrate the multiple sources of data that define medical problems, such as data that is derived from clinical protocols, laboratory measurements, and features extracted from signals and images.

Some recent research that approaches some of these broader goals include the work of [2, 9, 18, 24, 27]. One of the most promising techniques for improving both the flexibility and the accuracy of classification systems is to build systems that combine multiple classifiers [15]. The main idea behind a multiclassifier system is to average the hypotheses of a diverse group of classifiers, for instance, an ensemble of classifiers, in order to produce a better approximation to a true hypothesis [13]. In the last few years, a number of methods have been proposed in the literature for building multiclassifier systems.

In this chapter we provide an extensive review and evaluation of muliticlassifier techniques that can be reasonably used in data mining and in building practical decision support systems in healthcare. In section 2 we describe several methods for constructing ensembles of classifiers. We also describe state-of-the-art stand-alone classifiers (such as support vector machines, neural networks, and Gaussian process classifiers) that perform well on specific problems or, in the case of the Gaussian process classifier, in ensembles. Also included in this section is a description of some of the best feature transforms for extracting relevant information from noise. In section 3, we present our ensemble methods, using Rotation Forest, Input Decimated Ensemble, and rotation boosting that are based on variants of the feature transforms. In section 4, we describe several benchmark databases that provide a wide range of different kinds of medical data. In section 5, we apply our ensemble methods to these databases to illustrate the flexibility and accuracy of multiclassifier methods. Specifically, we compare several variants of the ensemble methods by varying the feature transform used to project the patterns. We also vary the method for selecting a set of training patterns for calculating the projections. We compare the results of our multiclassifier systems to the best stand-alone systems. Finally, in section 6, we summarize our results and make suggestions for further research.