Robust rules for prediction and description

Manuel Proenca, H.

Summary

Rules provide a simple form of storing and sharing information about the world. As humans, we use rules every day, such as the physician that diagnoses someone with flu, represented by "if a person has either a fever or sore throat (among others), then she has the flu.". Even though an individual rule can only describe simple events, several aggregated rules can describe more complex scenarios, such as the complete set of diagnostic rules employed by a physician.

Given their abundant use, it is no surprise that rule-based models were some of the first techniques used to equip computers with decision-making capabilities. In the beginning, humans entered rules directly into computer systems; however, with the availability of large amounts of data, the interest shifted to learning rules from data. For example, the records of a physician's diagnoses of patients who either have flu or not based on their symptoms can be used to learn that doctor's decision-making process. The use of rules spans many fields in computer science, and in this dissertation, we focus on rule-based models for machine learning and data mining. Machine learning focuses on learning from data the model that best predicts future (previously unseen) events. Data mining aims to find interesting patterns in the available data. Specifically, we are concerned with the research question: "How to learn robust and interpretable rule-based models from data for machine learning and data mining, and define their optimality?"

To answer such a question, we employ the Minimum Description Length (MDL) principle, which allows us to define the optimality of rule-based models for a particular dataset. Informally, the best model for a specific dataset is the simplest one that describes the data well. The specific model class we focus on is the rule list, i.e., an ordered set of rules that are interpreted sequentially. Nonetheless, finding an op-

timal model is computationally infeasible in most cases. Thus, we propose heuristic algorithms that find good models with some guarantees. We test our algorithms empirically to validate our approach and show that they achieve, in most cases, better or similar performance to the state-of-the-art in machine learning and data mining.