Chapter 4
Feature Based Rule Learner in Noisy Environment Using Neighbourhood Rough Set Model

Yang Liu
Xi'an Jiaotong University, China

Luyang Jiao
First Affiliated Hospital of Xinxiang Medical College, China

Guohua Bai
Blekinge Institute of Technology, Sweden

Boqin Feng
Xi'an Jiaotong University, China

ABSTRACT

From the perspective of cognitive informatics, cognition can be viewed as the acquisition of knowledge. In real-world applications, information systems usually contain some degree of noisy data. A new model proposed to deal with the hybrid-feature selection problem combines the neighbourhood approximation and variable precision rough set models. Then rule induction algorithm can learn from selected features in order to reduce the complexity of rule sets. Through proposed integration, the knowledge acquisition process becomes insensitive to the dimensionality of data with a pre-defined tolerance degree of noise and uncertainty for misclassification. When the authors apply the method to a Chinese diabetic diagnosis problem, the hybrid-attribute reduction method selected only five attributes from totally thirty-four measurements. Rule learner produced eight rules with average two attributes in the left part of an IF-THEN rule form, which is a manageable set of rules. The demonstrated experiment shows that the present approach is effective in handling real-world problems.

DOI: 10.4018/978-1-4666-0264-9.ch004

INTRODUCTION

Cognitive informatics is a study of information processing in humans, computers and in abstract (Wang, 2007a, 2007c; Yao, 2004). From the perspective of cognitive informatics, the cognition can be viewed as the acquisition of knowledge (Wang, 2007b). The study of knowledge discovery aims at building machine intelligence that facilitates human ways of thinking and understanding (Wang, 2009; Wang & Chiew, 2008). Many approaches have been implemented to improve the value of discovered knowledge intelligence (Chiew, 2002; Pazzani & Kibler, 1992; Wu, Bell, & David, 2003). The findings and in-depth understanding of knowledge discovery would have a significant impact on the advancement of cognitive informatics.

Knowledge acquisition system is the main implementation of knowledge discovery task. There have been many successful applications in real-world areas (Grzymala-Busse, 1988). However, complex application problems, such as reliable monitoring and diagnosis of diabetic patients, have emphasised the issue of knowledge acquisition and modelling. These problems are likely to present large number of features, not all of which will be essential for the task (Shen & Chouchoulas, 2002). Noisy and uncertain data cannot be ruled out. Furthermore, such applications typically require easily understood forms of knowledge that underlies data (Kurgan, 2004). Therefore, a method to allow automated generation of human comprehensible knowledge models with clear semantics in noisy environment is highly desirable (Kurgan & Musilek, 2006).

There are two most commonly used methods to generate expressive and human readable knowledge, i.e., decision trees and rule induction algorithms. Decision tree methods have drawn significant attention over the last several years, and incorporate various advanced speed, memory, and pruning optimization techniques (Quinlan, 1986). However, rule induction algorithms also exhibit a lot of desirable properties (Tsumoto, 2004). As reported by some problems (Cios & Kurgan, 2004), rule learner methods were found to outperform decision tree methods since the production of rule sets in expert system appears to be more human-comprehensible than decision trees (Michalski, 1983). In addition, rule set can be post-processed and analyzed in modality, which is very important when a decision maker needs to understand and validate the generated results, such as in medicine (Kurgan, Cios, & Dick, 2006).

LERS (Learning from examples based on rough sets) system is one of the most frequently used rough set based rule induction systems (Grzymala-Busse, 1992; Grzymala-Busse, Grzymala-Busse, & Hippe, 2001), which can handle inconsistencies using rough set theory, introduced by (Pawlak, 1982). Two algorithms LEM2 and MLEM2 are extensively used in LERS system since they perform better in applications in medicine (Grzymala-Busse, 2002). LERS system aims at optimizing the predictive performance on domain data. However, it does not meet the requirement of comprehensibility, such as the small size of rule set and the simple representation of rule forms (Liu, Bai, & Feng, 2008a). CompactLEM2 method was proposed to cope with large amounts of data and generate knowledge with simple representation (Liu, Bai, & Feng, 2008b). This method mainly measures the comprehensibility of rule set using two main factors: the number of rules, and the average length per rule. This strategy can reduce the complexity of rule set and extract compact knowledge, which in turn uses the small size of rule set and short rule forms to maintain high classification accuracy and imply useful information in datasets.

However, the real-world data sets are always not very well structured, and they tend to have high amounts of redundant attributes, which brings great difficulty in rule induction process (Guyon & Elisseeff, 2003). As nowadays thousands of attributes are stored in databases in some real-world applications, it is possible to remove irrelevant attributes to maintain the same classification