Chapter XIII
Re-Sampling Based Data Mining Using Rough Set Theory

Benjamin Griffiths
Cardiff University, UK

Malcolm J. Beynon
Cardiff University, UK

ABSTRACT

Predictive accuracy, as an estimation of a classifier's future performance, has been studied for at least seventy years. With the advent of the modern computer era, techniques that may have been previously impractical are now calculable within a reasonable time frame. Within this chapter, three techniques of resampling, namely, leave-one-out, k-fold cross validation and bootstrapping; are investigated as methods of error rate estimation with application to variable precision rough set theory (VPRS). A prototype expert system is utilised to explore the nature of each resampling technique when VPRS is applied to an example dataset. The software produces a series of graphs and descriptive statistics, which are used to illustrate the characteristics of each technique with regards to VPRS, and comparisons are drawn between the results.

INTRODUCTION

The success of data mining is dependent on the appropriateness and practicality of the mathematical analysis techniques employed. Whether it is based on statistical or symbolic machine learning, an analyst is interested primarily in the final results that they can analyse and interpret. Those theorists interested in researching data mining are aware that one-off analyses using different techniques may produce varying results, including that of the often definitive statistic of predic-

tive accuracy (as for example in classification problems). More pertinently, these results may highlight further incongruousness such as different characteristic attributes in a dataset viewed as important. For the analyst they are then in a quandary, since they desire confidence in the concomitant interpretation.

One option to mitigate the limited confidence that can be inherent with a one-off analysis, with any technique, is through the use of resampling. For the last 70 years, since the early work on error rate estimation (e.g., Larson, 1931), there has consistently been research undertaken on the benefits of resampling based analysis (e.g., Efron, 1982; Braga-Neto & Dougherty, 2005). Whilst research has considered the statistical consequences of such resampling and related analysis (e.g., Shao & Tu, 1995), how nascent techniques, in particular those based on symbolic machine learning, fully utilize such approaches requires elucidation.

Two questions to consider are; whether accustomed parameters used in resampling are generally appropriate to all data mining techniques, and beyond classification accuracy results are resampling approaches appropriate to elucidate insights such as attribute importance. A relevant example of such an overlapping philosophy is with the work of Leo Breiman, who has undertaken extensive research on the issue of resampling (e.g., Breiman, 1996), but also advocates the need to develop new data mining techniques (e.g., Breiman, 2001).

This chapter demonstrates these questions and the role of resampling in data mining, with emphasis on the results from rough set theory (RST). As a nascent symbolic machine learning technique (Pawlak, 1982), its popularity is a direct consequence of its operational processes, which adhere most closely to the notions of knowledge discovery and data mining (Li & Wang, 2004). Characteristics like this contribute to the adage given in Dunstch and Gediga (1997, p. 594), that underlying the RST philosophy is: "*Let the data speak for itself*."

However, its novel set theoretical structure brings its own concerns when considered within a resampling environment. Consequently, this chapter offers insights into the resampling issues that may affect practical analysis when employing nascent techniques, such as RST.

The specific data mining technique utilised here is *variable precision rough set theory* (VPRS, Beynon, 2001; Ziarko, 1993), one of a number of developments on the original RST. Other developments of RST include; dominance-based rough sets (Greco, Matarazzo, & Słowiński, 2004), fuzzy rough sets, (Greco, Inuiguchi, & Słowiński, 2006) and probabilistic rough sets (Ziarko, 2005). A literal presentation of the diversity of work on RST can be viewed in the annual volumes of the transactions on rough sets (most recent year 2005). The utilisation of VPRS is without loss of generality to developments such as those referenced; its relative simplicity allows the non-proficient reader the opportunity to follow fully the details presented.

The result from a VPRS analysis is a group of "*if .. then ..*" decision rules, which classify objects to decision classes based on their condition attribute values. VPRS also allows for the misclassification of some objects used in the construction of the decision rules, unlike in RST. One relevant issue is the intent within VPRS (and RST in general) for data reduction and feature selection (Jensen & Shen, 2005), with subsets of condition attributes identified that perform the same role as all the condition attributes in a considered dataset (termed β-reducts in VPRS). These β-reducts are identified prior to the construction of its respective group of decision rules, rather than during their construction as in techniques like decision trees, an issue pertinent when resampling is employed (Breiman, 2001).

A number of resampling approaches are considered here, namely; "Leave-one-out" (Weiss & Kulikowski, 1991), *k*-fold cross validation (Braga-Neto & Dougherty, 2004) and bootstrapping (Chennick, 1999). The VPRS results presented