Information Veins and Resampling with Rough Set Theory

Benjamin Griffiths
Cardiff University, UK

Malcolm J. Beynon
Cardiff University, UK,

INTRODUCTION

Rough Set Theory (RST), since its introduction in Pawlak (1982), continues to develop as an effective tool in data mining. Within a set theoretical structure, its remit is closely concerned with the classification of objects to decision attribute values, based on their description by a number of condition attributes. With regards to RST, this classification is through the construction of '*if .. then ..*' decision rules. The development of RST has been in many directions, amongst the earliest was with the allowance for miss-classification in the constructed decision rules, namely the Variable Precision Rough Sets model (VPRS) (Ziarko, 1993), the recent references for this include; Beynon (2001), Mi et al. (2004), and Ślęzak and Ziarko (2005). Further developments of RST have included; its operation within a fuzzy environment (Greco et al., 2006), and using a dominance relation based approach (Greco et al., 2004).

The regular major international conferences of 'International Conference on Rough Sets and Current Trends in Computing' (RSCTC, 2004) and 'International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing' (RSFDGrC, 2005) continue to include RST research covering the varying directions of its development. This is true also for the associated book series entitled 'Transactions on Rough Sets' (Peters and Skowron, 2005), which further includes doctoral theses on this subject. What is true, is that RST is still evolving, with the eclectic attitude to its development meaning that the definitive concomitant RST data mining techniques are still to be realised. Grzymala-Busse and Ziarko (2000), in a defence of RST, discussed a number of points relevant to data mining, and also made comparisons between RST and other techniques.

Within the area of data mining and the desire to identify relationships between condition attributes, the effectiveness of RST is particularly pertinent due to the inherent intent within RST type methodologies for data reduction and feature selection (Jensen and Shen, 2005). That is, subsets of condition attributes identified that perform the same role as all the condition attributes in a considered data set (termed β-reducts in VPRS, see later). Chen (2001) addresses this, when discussing the original RST, they state it follows a reductionist approach and is lenient to inconsistent data (contradicting condition attributes - one aspect of underlying uncertainty). This encyclopaedia article describes and demonstrates the practical application of a RST type methodology in data mining, namely VPRS, using nascent software initially described in Griffiths and Beynon (2005). The use of VPRS, through its relative simplistic structure, outlines many of the rudiments of RST based methodologies.

The software utilised is oriented towards 'hands on' data mining, with graphs presented that clearly elucidate 'veins' of possible information identified from β-reducts, over different allowed levels of miss-classification associated with the constructed decision rules (Beynon and Griffiths, 2004). Further findings are briefly reported when undertaking VPRS in a resampling environment, with leave-one-out and bootstrapping approaches adopted (Wisnowski et al., 2003). The importance of these results is in the identification of the more influential condition attributes, pertinent to accruing the most effective data mining results.

BACKGROUND

VPRS development on RST is briefly described here (see Ziarko, 1993; Beynon, 2001). It offers the allowance for miss-classification of objects in the constructed decision rules (determined by a β value over its allowed domain, see later), as well as correct classification and

the non-classification of objects (the β value infers a level of certainty in the model). This is one of a number of directions of development in RST based research. By way of example, the Bayesian rough set model moves away from the requirement for a particular β value (Ślęzak and Ziarko, 2005), instead it considers an appropriate certainty gain expression (using Bayesian reasoning). The relative simplicity of VPRS offers the reader the opportunity to perceive the appropriateness of RST based approaches to data mining.

Central to VPRS (and RST) is the information system (termed here as the data set), which contains a universe of objects $U(o_1, o_2, \ldots)$, each characterised by a set condition attributes $C(c_1, c_2, \ldots)$ and classified to a set of decision attributes $D(d_1, d_2, \ldots)$. Through the indiscernibility of objects based on C and D, respective condition and decision classes of objects are found. For a defined proportionality value β, the β-positive region corresponds to the union of the set of condition classes (using a subset of condition attributes P), with conditional probabilities of allocation to a set of objects Z (using a decision class $Z \in E(D)$), which are at least equal to β. More formally:

$$\beta\text{-positive region of the set } Z \subseteq U \text{ and } P \subseteq C : POS_P^{\beta}(Z) = \bigcup_{\Pr(Z \mid X_i) \geq \beta} \{X_i \in E(P)\},$$

where β is defined here to lie between 0.5 and 1 (Beynon, 2001), and contributes to the context of a *majority inclusion* relation. That is, those condition classes in a β-positive region, $POS_P^{\beta}(Z)$, each have a majority of objects associated with the decision class $Z \in E(D)$. The numbers of objects included in the condition classes that are contained in the respective β-positive regions for each of the decision classes, subject to the defined β value, make up a measure of the *quality of classification*, denoted $\gamma^{\beta}(P, D)$, and given by:

$$\gamma^{\beta}(P, D) = \frac{\text{card}(\bigcup_{Z \in E(D)} POS_P^{\beta}(Z))}{\text{card}(U)},$$

where $P \subseteq C$. The $\gamma^{\beta}(P, D)$ measure with the β value means that for the objects in a data set, a VPRS analysis may define them in one of three states; not classified, correctly classified and miss-classified. Associated with this is the $\beta_{\min}$ value, the lowest of the (largest) proportion values of β that allowed the set of condition classes to be in the β-positive regions constructed. That is, a β value above this upper bound would imply at least one of the contained condition classes would then not be given a classification.

VPRS further applies these defined terms by seeking subsets of condition attributes (termed β-*reducts*), capable of explaining the associations given by the whole set of condition attributes, subject to the majority inclusion relation (using a β value). Within data mining, the notion of a β-reduct is directly associated with the study of data reduction and feature selection (Jensen and Shen, 2005). Ziarko (1993) states that a β-reduct (R) of the set of conditional attributes C, with respect to a set of decision attributes D, is:

i) A subset R of C that offers the same quality of classification, subject to the β value, as the whole set of condition attributes.

ii) No proper subset of R has the same quality of the classification as R, subject to the associated β value.

An identified β-reduct is then used to construct the decision rules, following the approach utilised in Beynon (2001). In summary, for the subset of attribute values that define the condition classes associated with a decision class, the values that discern them from the others are identified. These are called prime implicants and form the condition parts of the constructed decision rules (further reduction in the prime implicants also possible).

Main Focus

When large data sets are considered, the identification of β-reducts and adopted balance between classification/miss-classification of objects infers small veins of relevant information are available within the whole β domain of (0.5, 1]. This is definitive of data mining and is demonstrated here using the VPRS software introduced in Griffiths and Beynon (2005). A large European bank data set is utilised to demonstrate the characteristics associated with RST in data mining (using VPRS).

The Bank Financial Strength Rating (BFSR) introduced by Moody's rating agency is considered (Moody's, 2004), which represent their opinion of a bank's intrinsic safety and soundness (see Poon et al., 1999). Thirteen BFSR levels exist, but here a more general grouping of ratings of (A or B), C and (D or E) are considered (internally labelled 0 to 2). This article considers an exhaustive set of 309 European banks for which a BFSR rating has been assigned to them. The numbers of banks assigned a specific rating is; (A or B)