Rough Set Theory

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 classification problems and decision support. In the majority of applications using RST based methodologies, there is the construction of '*if .. then ..*' decision rules that are used to describe the results from an analysis. The variation of applications in management and decision making, using RST, recently includes discovering the operating rules of a Sicilian irrigation purpose reservoir (Barbagallo, Consoli, Pappalardo, Greco, & Zimbone, 2006), feature selection in customer relationship management (Tseng & Huang, 2007) and decisions that insurance companies make to satisfy customers' needs (Shyng, Wang, Tzeng, & Wu, 2007).

As a nascent symbolic machine learning technique, the popularity of RST is a direct consequence of its set theoretical operational processes, mitigating inhibiting issues associated with traditional techniques, such as within-group probability distribution assumptions (Beynon & Peel, 2001). Instead, the rudiments of the original RST are based on an indiscernibility relation, whereby objects are grouped into certain equivalence classes and inference taken from these groups. Characteristics like this mean that decision support will be built upon the underlying RST philosophy of "*Let the data speak for itself*" (Dunstch & Gediga, 1997). Recently, RST was viewed as being of fundamental importance in artificial intelligence and cognitive sciences, including decision analysis and decision support systems (Tseng & Huang, 2007).

One of the first developments on RST was through the variable precision rough sets model ($VPRS_{\beta}$), which allows a level of mis-classification to exist in the classification of objects, resulting in probabilistic rules (see Ziarko, 1993; Beynon, 2001; Li and Wang, 2004). $VPRS_{\beta}$ has specifically been applied as a potential decision support system with the UK Monopolies and Mergers Commission (Beynon & Driffield, 2005), predicting bank credit ratings (Griffiths & Beynon, 2005) and diffusion of medicaid home care programs (Kitchener, Beynon, & Harrington, 2004).

Further developments of RST include extended variable precision rough sets ($VPRS_{l,u}$), which infers asymmetric bounds on the possible classification and mis-classification of objects (Katzberg & Ziarko, 1996), dominance-based rough sets, which bases their approach around a dominance relation (Greco, Matarazzo, & Słowiński, 2004), fuzzy rough sets, which allows the grade of membership of objects to constructed sets (Greco, Inuiguchi, & Słowiński, 2006), and probabilistic bayesian rough sets model that considers an appropriate certainty gain function (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 2006), also the annual conferences dedicated to RST and its developments (see for example, RSCTC, 2004). In this article, the theory underlying $VPRS_{l,u}$ is described, with its special case of $VPRS_{\beta}$ used in an example analysis. The utilisation of $VPRS_{l,u}$, and $VPRS_{\beta}$, is without loss of generality to other developments such as those referenced, its relative simplicity allows the non-proficient reader the opportunity to fully follow the details presented.

BACKGROUND

The background to the whole range of RST based methodologies is beyond the scope of a single book chapter, here one line of development is described, and is illustrative of RST and its evolution. Moreover, the original RST, $VPRS_{\beta,}$ and $VPRS_{l,u}$ methodologies are discussed and how they are related further exposited (see Beynon, 2003). Central to the RST associated methodologies is the information system, in the form of a decision table.

A decision tables is made up of a set of objects (U), each characterized by a set of categorical condition attributes (C) and classified by a set of categorical decision attributes (D). A value denoting the nature of

an attribute to an object is called a *descriptor*. From C and D, certain equivalence classes (condition and decision) are constructed through the utilisation of an indiscernibility relation (unlike, for example, the use of the dominance relation in dominance based rough sets, see Greco et al., 2004). Using an indiscernibility relation, a condition class contains objects that have the same categorical condition attribute values (similar for a decision class). The use of categorical data here means that a level of data discretisation is necessary if continuous data are present (see Beynon & Peel, 2001).

Within the original RST, the decision rules constructed are deterministic, meaning they do not allow for a level of mis-classification of objects to a decision class. That is, for a condition class given a classification, the contained objects are all classified to the same decision class. RST was developed to allow a level of mis-classification, by the inclusion of the $\boldsymbol{\beta}$-threshold value of necessary majority inclusion in the condition classes given a classification (to the same decision class), called the variable precision rough sets model ($\text{VPRS}_{\boldsymbol{\beta}}$). This was further developed with the extended variable precision rough sets model ($\text{VPRS}_{l,u}$), where asymmetric bounds l and u are used to control objects given and not given a classification to a decision class. The utilisation of these bounds, l and u, enable the construction of certain set approximations (regions), including the u-positive region $POS_P^u(Z)$ of a set, defined by:

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

where $E(\cdot)$ denotes an equivalence class ($E(P)$—condition classes from the set of condition attributes P), and u reflects the least acceptable degree of the conditional probability $\Pr(Z \mid X_i)$ of objects in the condition class X_i, to include X_i in the u-positive region for Z. An analogous l-negative region $NEG_P^l(Z)$ is given by:

l-negative region of the set $Z \subseteq U$ and $P \subseteq C$: $NEG_P^l(Z) = \bigcup_{\Pr(Z \mid X_i) \leq l} \{X_i \in E(P)\}$,

where l reflects the largest acceptable degree of the conditional probability $\Pr(Z \mid X_i)$ to include the condition class X_i in the l-negative region. A (l, u)-boundary region $BND_P^{l,u}(Z)$ represents those condition classes that cannot be classified to $Z \subseteq U$ with sufficiently high confidence (not greater than u) and cannot be excluded from the classification to Z (not less than l), which is given by:

(l, u)-boundary region of the set $Z \subseteq U$ and $P \subseteq C$: $BND_P^{l,u}(Z) = \bigcup_{l < \Pr(Z \mid X_i) < u} \{X_i \in E(P)\}$.

From these definitions, the objects contained in a decision table can be partitioned into one of the three defined approximation regions. Further, to these approximation regions certain measures can be constructed with respect to the objects in the decision table.

The one measure considered here is a direct progression from RST and $\text{VPRS}_{\boldsymbol{\beta}}$, and concerns the *quality of classification* (*QoC*). This measure relates to the proportion of objects in a decision table, which are included in the associated u-positive regions. That is, the proportion of objects that are classified to single decision classes. With $\text{VPRS}_{l,u}$, for all the objects in the set $Z \subseteq U$, the quality of classification (l, u)-*QoC* ($\boldsymbol{\gamma}^{l,u}(P, D)$) is given by:

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

where $P \subseteq C$. It is noted, the measure $\boldsymbol{\gamma}^{l,u}(P, D)$ is dependent only on the u boundary value, because it is concerned with objects that are classified to a single decision class. The $\boldsymbol{\gamma}^{l,u}(P, D)$ measure with the l and u values 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 mis-classified.

The approximation regions describing $\text{VPRS}_{l,u}$ are general definitions that also describe the previously introduced RST and $\text{VPRS}_{\boldsymbol{\beta}}$ methodologies. For the original RST, its approximation regions are described by the choice of l and u, when l=0 and u=1, subsequently defined POS_P^1, NEG_P^0, and $BND_P^{0,1}$. For $\text{VPRS}_{\boldsymbol{\beta}}$, the choice of l and u is with respect to the single value $\boldsymbol{\beta}$ and is when $\boldsymbol{\beta} = u = 1 - l \in (0.5, 1]$, with the approximation regions defined $POS_P^{\boldsymbol{\beta}}$, $NEG_P^{1-\boldsymbol{\beta}}$, and $BND_P^{1-\boldsymbol{\beta},\boldsymbol{\beta}}$.

The measures described here highlight the generality implicit in $\text{VPRS}_{l,u}$. That is, the allowance for the l and u values to take any values over the [0, 1] domain, subject to $l < u$. Generally, the marginal effect on the described measures of object classification due to changes in either of the l and u values is difficult to perceive. This lack of perception was identified in Beynon (2003), who introduced the (l, u)-graph, see Figure 1.