Introduction and Elucidation of the Quality of Sagacity in the Extended Variable Precision Rough Sets Model

Malcolm J. Beynon [1]

Cardiff Business School, Cardiff University, Colum Drive, Cardiff, CF10 3EU, Wales, UK

Abstract

This paper introduces the quality of sagacity measure in the extended variable precision rough sets model - $\text{VPRS}_{l,u}$. The need for this measure is a direct consequence of the use of the associated l and u values. Moreover, different levels of miss-classification are allowed in the classification of objects to a decision class or its compliment. This measure attempts to take this into account, by acknowledging the classification of an object to the compliment of a decision class, as an appropriate (if not optimum) classification of that object. A consequence of the analysis is a discussion of the notion of open and closed worlds in $\text{VPRS}_{l,u}$.

1 Introduction

The extended variable precision rough sets model ($\text{VPRS}_{l,u}$) developed by Katzberg and Ziarko [3] includes a general allowance for levels of miss-classification in an objects classification. The associated l and u values define the construction of certain set approximation regions which define the possible classification of objects. Importantly, the l and u values introduce different levels of allowed miss-classification of an object to a decision class or its compliment.

The degree of dependency measure [3] acknowledges this aspect through measures on the individual decision classes. A quality of classification measure concentrates on the objects classified to a specific single decision class. To include the classification of objects to the compliment of a decision class, while less warranted than the actual classification of objects to a decision class it does discern some classification knowledge on objects otherwise not given a classification. Through the utilisation of a (l, u)-graph [1] the partition of objects to each of the two different types of classification is elucidated.

[1] Email: BeynonMJ@cardiff.ac.uk

An initial measure augmenting the possible classification of objects is introduced, defined the quality of sagacity.[2] Interpreting the classification of an object to the compliment of a decision class in two ways enables an informal discussion on the notion of open and closed world in $\text{VPRS}_{l,u}$.

2 Brief Description of Extended VPRS

Within a decision table there exists a set objects (U) each characterized and classified by sets of condition (C) and decision (D) attributes respectively $\text{VPRS}_{l,u}$ allows for probabilistic classification by utilising two control parameters, defined lower (l) and upper (u) limits respectively, constrained by $0 \leqslant l < u \leqslant 1$. From C and D, certain condition and decision equivalence classes $E(C)$ and $E(D)$ respectively are constructed. With $Z \subseteq U$ and $P \subseteq C$ three approximation regions are defined, firstly the u-positive and l-negative regions:

$$POS_u(Z) = \bigcup \{X_i \in E(P) : \Pr(Z/X_i) \geqslant u\},$$

$$NEG_l(Z) = \bigcup \{X_i \in E(P) : \Pr(Z/X_i) \leqslant l\},$$

where in each region $\Pr(Z \,/\, X_i)$ is a conditional probability estimate of Z given $X_i \in E(P)$. The $POS_u(Z)$ and $NEG_l(Z)$ regions represent the acceptability of the membership of X_i as being likely or unlikely to belong to Z respectively (subject to l and u). Where there is no acceptable likeliness, then the X_i is a member of the (l,u)-boundary region, defined by

$$BNR_{l,u}(Z) = \bigcup \{X_i \in E(P) : l < \Pr(Z/X_i) < u\}.$$

That is, $X_i \in BNR_{l,u}(Z)$ cannot be classified to Z or the compliment of Z with an acceptable error rate. Within $\text{VPRS}_{l,u}$, the notion of the compliment of Z (defined $U - Z$) is important since the utilisation of l and u values means the possible classification of an object to Z or $U - Z$ is with respect to different levelsof miss-classification. The acknowledgement of this difference in the levels of classification (use of l and u) implies that classifying to the compliment of Z is itself some form of classification. It follows, the proportion of the objects classified to a decision class or its compliment can be represented by the $\gamma^{l,u}(P,D)$ and $\gamma_{\text{C}}^{l,u}(P,D)$ values respectively, and given by

$$\gamma^{l,u}(P,D) = \frac{\text{card}(\bigcup_{X_j \in E(P)} \{X_i | \exists\, j \;\; X_i \in POS_u(D_j)\})}{\text{card}(U)},$$

and

[2] The term sagacity implies wise or good judgement and signifies the taking into account of more than just classification to a single decision class, but also to its compliment.

$$\gamma_{\mathrm{C}}^{l,u}(P,D) = \frac{\mathrm{card}(\bigcup_{X_j \in E(P)} \{X_i | \exists_j\, X_i \in NEG_l(D_j) \text{ and } \forall_j\, X_i \notin POS_u(D_j)\})}{\mathrm{card}(U)}.$$

The $\gamma^{l,u}(P,D)$ expression is analogous to the quality of classification expression in the variable precision rough set model - VPRS_β with symmetrical bound β [4]. That is, the proportion of objects of U in those condition classes which are contained in a $POS_u(D_j)$ region. Whereas $\gamma_{\mathrm{C}}^{l,u}(P,D)$ is the proportion of objects of U in condition classes which are not in any $POS_u(D_j)$ regions, but in a $NEG_l(D_j)$ region.

Allowing the classification of objects to the compliment of a decision class then the quality of sagacity in the $\mathrm{VPRS}_{l,u}$ case is next defined. With respect to all the objects in the set U, the $(l,\ u)$-quality of sagacity (l,u)-QoS $(\sigma^{l,u}(P,D))$ is given by

$$\sigma^{l,u}(P,D) = 1 - \frac{\mathrm{card}(\bigcap_{D_j \in E(D)} BNR_{l,u}(D_j))}{\mathrm{card}(U)}.$$

The $\sigma^{l,u}(P,D)$ measure represents the proportion of objects (possibly subject to a level of miss-classification) which are classified to a decision class or the complement of a decision class. That is, subject to the l and u values $\sigma^{l,u}(P,D)$ does not consider those objects, which cannot be included in any of the $POS_u(D_j)$ and $NEG_l(D_j)$ regions. It is noted a connection between these expressions is given by $\gamma^{l,u}(P,\ D) + \gamma_{\mathrm{C}}^{l,u}(P,D) = \sigma^{l,u}(P,D)$.

3 Description of data

In this paper the wine data set is utilised, which consists of different wines derived from three different cultivators (making up the decision classes D_1, D_2 and D_3). Three (out of 13) condition attributes (see Table 1) and 40 (out of 178) different wines (see Appendix A) are considered. Since all the attributes are continuous in nature, for a $\mathrm{VPRS}_{l,u}$ analysis, they need to be discretised into intervals. Table 1 shows the results of the discretisation, found using the minimum-entropy method [2]. The minimum-entropy method requires a decision on the number of intervals to discretise each continuous attribute into; in this case two intervals were used (labelled 1 and 2).

In Table 1, the three condition attributes are each discretised into two intervals, also shown is the number of wines in each interval, now in categorical form. The categorical descriptor values enable the set of objects to be included in a number of condition and decision classes. In this case there are five condition classes X_1, X_2, X_3, X_4 and X_5 which include all the 40 wines (see Table 2).

In Table 2, the descriptor values identifying each object to a condition

Table 1
Description of Condition attributes

Condition Attribute	Interval	
	1	2
c_1 - Malic acid	[0.8900, 2.6550], 25	[2.6550, 4.6100], 15
c_2 - Ash	[1.7000, 2.0400], 36	[2.0400, 2.8000], 4
c_3 - Magnesium	[11.2000, 19.800], 23	[19.8000, 25.5000], 17

class is given. The number of objects in each condition class are reported as well as the number (and proportion) of these objects which are in each of the three decision classes. The values at the bottom of each D_j column indicate the number of objects (and proportions) in each decision class.

Table 2
Condition and decision classes in wine problem

Condition classes	$D_1 - 1$	$D_2 - 2$	$D_3 - 3$
X_1 - $\{c_1 = 1, c_2 = 2, c_3 = 1\}$, 12	4 (0.3333)	7 (0.5833)	1 (0.0833)
X_2- $\{c_1 = 1, c_2 = 2, c_3 = 2\}$, 9	7 (0.7778)	1 (0.1111)	1 (0.1111)
X_3- $\{c_1 = 2, c_2 = 2, c_3 = 2\}$, 8	2 (0.2500)	0 (0.0000)	6 (0.7500)
X_4- $\{c_1 = 2, c_2 = 2, c_3 = 1\}$, 7	1 (0.1429)	1 (0.1429)	5 (0.7142)
X_5- $\{c_1 = 1, c_2 = 1, c_3 = 1\}$, 4	0 (0.0000)	4 (1.0000)	0 (0.0000)
Decision classes	14 (0.3500)	13 (0.3250)	13 (0.3250)

4 (l, u)-graphs describing (l, u)-QoS in $\mathbf{VPRS}_{l,u}$

In this section (l, u)-graphs introduced in Beynon [1] see Fig.1, are produced to aid in the elucidation of the descriptive measure (l, u)-QoS.

In Fig.1, the general (l, u)-graph is presented, and shows the domain of the (l, u)-space is an equilateral triangle, its shape is governed by the constraint $0 \leqslant l < u \leqslant 1$. A general point is described by (l, u), for example in Fig.1 two choices of l and u values are shown, in the case when $l = 0.1$, $u = 0.8$ and $l = 0.6$, $u = 0.7$. In the subsequent (l, u)-graphs presented, regions of the graph are identified which have the same level of the measure being considered. Before the (l, u)-QoS graphs are exposited, the partition of objects classified to a decision class or its compliment are described by the ordered list $[\gamma^{l,u}(P, D), \gamma^{l,u}_{\mathrm{C}}(P, D)]$ for the wine data set, and are reported in Fig.2 in the form of regions of the (l, u)-graph with the same $[\gamma^{l,u}(C, D), \gamma^{l,u}_{\mathrm{C}}(P, D)]$ pair of values (with the ordered lists showing actual numbers of objects not proportions - for proportions need divide each value by 40).

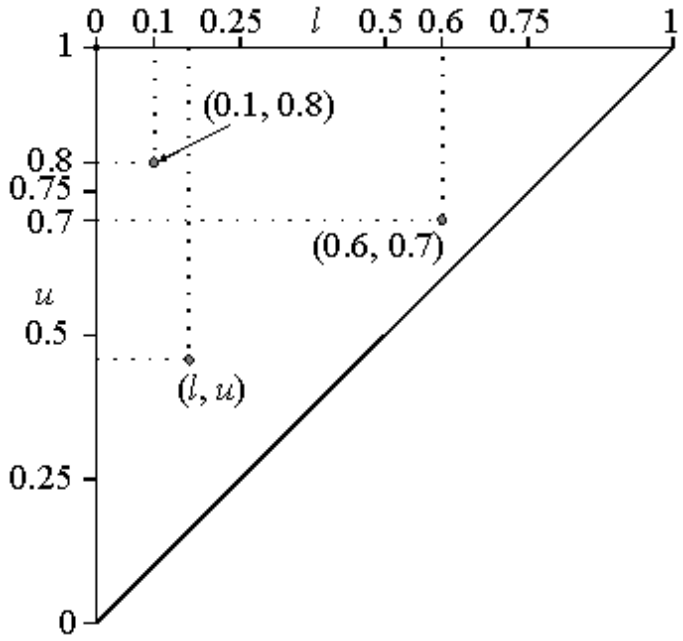

Fig. 1. General (l, u)-graph

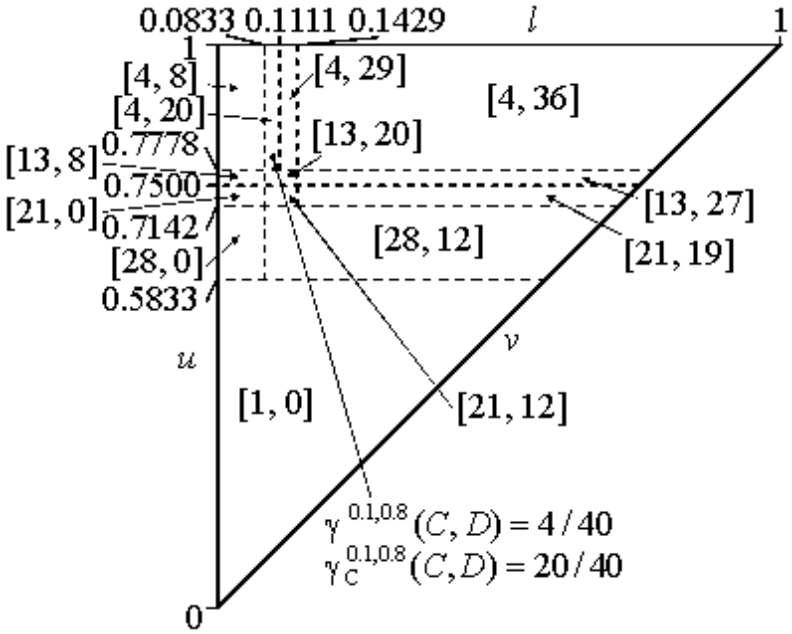

Fig. 2. (l, u)-graph showing regions of different $[\gamma^{l,u}(C,D),\ \gamma_{\mathrm{C}}^{l,u}(C,D)]$ values

T o illustrate the constructionof the $\gamma^{l,u}(C,D)$ and $\gamma_{\mathrm{C}}^{l,u}(C,D)$ values identified in regions of the (l, u)-graphs the specific case of $l = 0.1$ and $u = 0.8$ is considered, and with respect to Table 2 are given b y

$$\gamma^{0.1,0.8}(C,D) = \frac{\mathrm{card}(\bigcup_{X_j \in E(C)} \{X_i | \exists_j \text{ such that } X_i \in POS_{0.8}(D_j)\})}{\mathrm{card}(U)} =$$

$$= \frac{\text{card}(\{X_5\})}{40} = \frac{4}{40}$$

and

$$\gamma_{\text{C}}^{0.1,0.8}(C,D) = \frac{\text{card}(\bigcup_{X_j \in E(C)} \{X_i | \exists_j X_i \in NEG_{0.1}(D_j) \text{ and } \forall_j X_i \notin POS_{0.8}(D_j)\})}{\text{card}(U)}$$

$$= \frac{\text{card}(\bigcup_{X_j \in E(C)} \{X_i | X_i \in \{X_1, X_3, X_5\} \text{ and } X_i \notin \{X_5\}\})}{40}$$

$$= \frac{\text{card}(\{X_1, X_3\}\ )}{40} = \frac{20}{40}$$

Hence (ignoring the divide by 40 part) we get the ordered list [4, 20] as shown in Fig.2. This interprets to; of the 40 wines, based on $l = 0.1$ and $u = 0.8$ then 4 wines are able to be classified to a single decision class and 20 only able to be classified to the compliment of single decision classes. To combine these levels of classification into a single measure then the (l,u)-QoS measure is considered. With respect to the (l, u)-QoS $\sigma^{l,u}(C,D)$, Fig.3 shows the regions of the (l, u)-QoS graph with different $\sigma^{l,u}(C, D)$ values.

In Fig. 3 the (l, u)-QoS graph shows a large region with $\sigma^{l,u}(C,D) = 1$, which signifies all objects are classified to a decision class or the compliment of a decision class. Of particular interest is where $\sigma^{l,u}(C,D) < 1$, here the (l,u)-range of the associated $BNR_{l,u}(\cdot)$ regions is large - l and u mostly towards 0 and 1 respectively . Indeed for the region $\sigma^{l,u}(C,D) < 1$, it satisfies $u - l > 0.5$ (with $u - l = 0.5$ when $u = 0.5833$ and $l = 0.0833$). To illustrate the construction of the (l,u)-QoS graphs, the calculation of $\sigma^{0.1,0.8}(C,D)$ with $l = 0.1$ and $u = 0.8$ (shown in Fig.3) is next given

$$\sigma^{0.1,0.8}(C,D) = 1 - \frac{\text{card}(\bigcap_{D_j \in E(D)} BNR_{0.1,0.8}(D_j))}{\text{card}(U)}$$

$$= 1 - \frac{\text{card}(BNR_{0.1,0.8}(D_1) \cap BNR_{0.1,0.8}(D_2) \cap BNR_{0.1,0.8}(D_3))}{\text{card}(U)}$$

$$= 1 - \frac{\text{card}((\cup\{X_1, X_2, X_3, X_4\}) \cap (\cup\{X_1, X_2, X_4\}) \cap (\cup\{X_2, X_3, X_4\}))}{\text{card}(U)}$$

$$= 1 - \frac{\text{card}(\cup\{X_2, X_4\})}{\text{card}(U)} = 1 - \frac{16}{40} = \frac{24}{40}$$

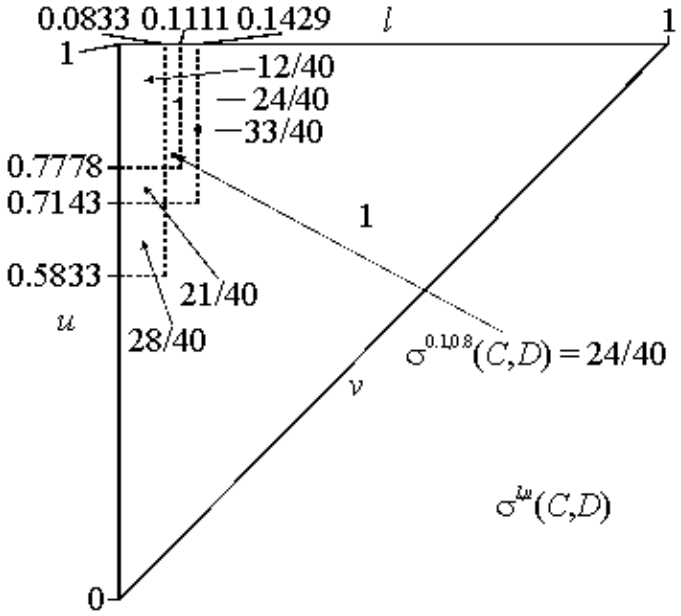

Fig. 3. (l, u)-QoS graph with $C = \{c_1, c_2, c_3\}$

This calculation also confirms $\gamma^{0.1,0.8}(C, D)+\gamma_{\mathrm{C}}^{0.1,0.8}(C, D) = \sigma^{0.1,0.8}(C, D)$ (as does comparison of Fig. 2 with Fig. 3 in general). Similar (l, u)-graphs can be constructed for the (l, u)-QoS measure for subsets of the condition attributes C. That is, the $\sigma^{l,u}(P, D)$ values for $P \subset C$, with three condition attributes considered there exist six proper subsets to consider, see Fig. 4.

Following the approach in Beynon [1], in each of the six (l, u)-QoS graphs in Fig. 4, the shaded regions show the subdomain of the (l, u) domain space for which the subset of condition attributes is a (l, u)-reduct for this problem. Understandably the region of unshaded area is near the top left corner of the (l, u)-graph for which each $BNR_{l,u}(\cdot)$ has a relatively large (l, u)-range. That is, the l and u values associated with this area of the (l, u)-graph are near their extreme values 0 and 1 respectively. The proportion of area ($PoA_{QoS}(\cdot)$) and proportion of discernibility ($PoD_{QoS}(\cdot)$) measures [1] associated with the (l, u)-QoS measure are next identified for each of the (l, u)-QoS graphs in Fig. 4, see Table 3.

Table 3
Description of Condition attributes

P	$\{c_1\}$	$\{c_2\}$	$\{c_3\}$	$\{c_1, c_2\}$	$\{c_1,c_3\}$	$\{c_2,c_3\}$
$PoA_{QoS}(\cdot)$	0.8809	0.6945	0.7921	0.8821	0.9306	0.6953
$PoD_{QoS}(\cdot)$	0.9168	0.6945	0.7921	0.9093	0.9150	0.6953

In Table 3 the $PoA_{QoS}(\cdot)$ and $PoD_{QoS}(\cdot)$ values differ for different subsets of the condition attributes. These values can be used to identify a possible single (l, u)-reduct. With three condition attributes, (l, u)-reducts of different sizes can be identified. For each different size of possible (l, u)-reduct, from

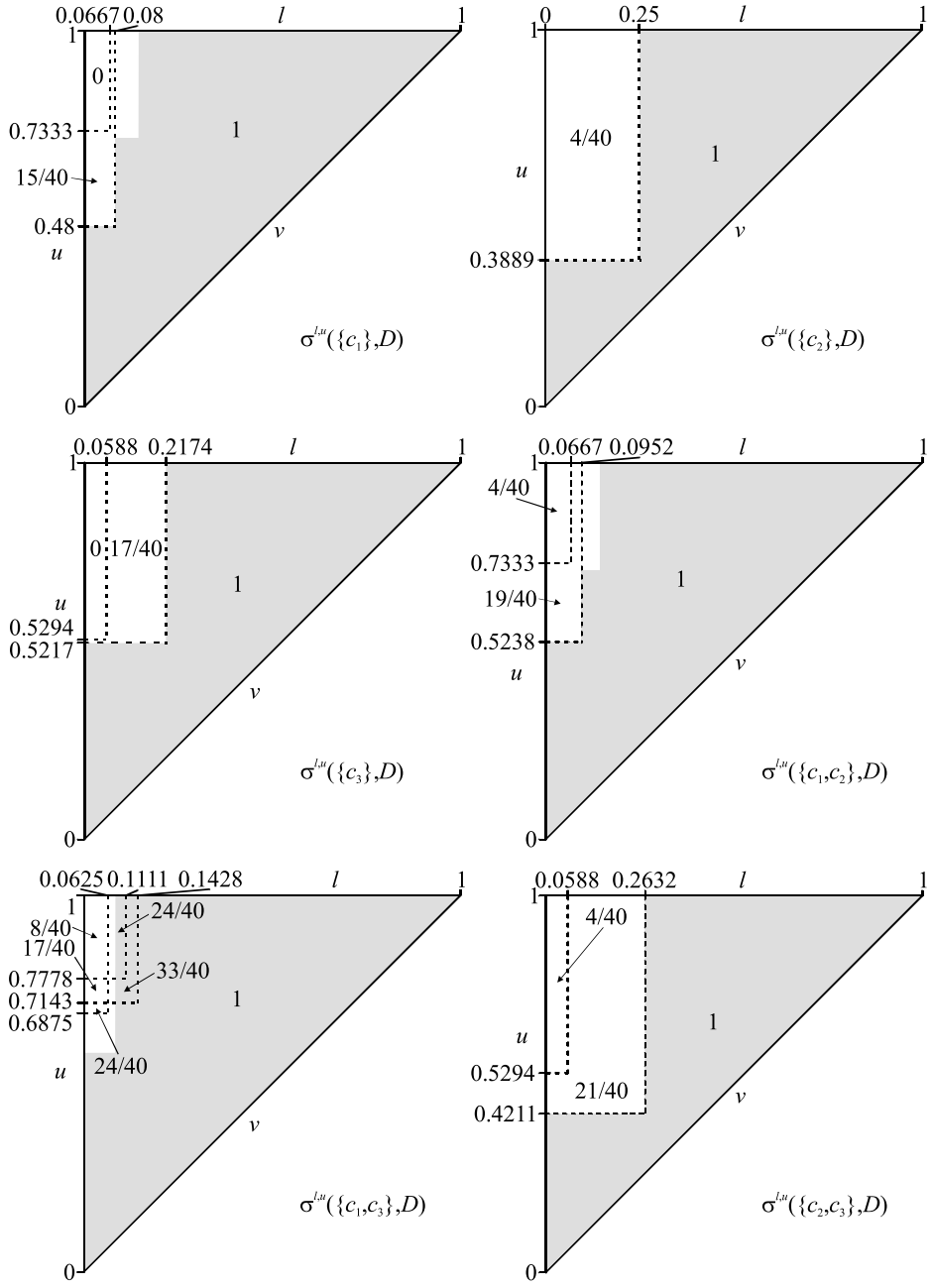

Fig. 4. $(l,\ u)$-QoS graphs for $\sigma^{l,u}(P,\ D)$ with $P \subset C = \{c_1,\ c_2,\ c_3\}$

T able3 the subsets $\{c_1\}$ and $\{c_1,c_3\}$ hav e the largest $PoA_{QoS}(\cdot)$ values from amongst those subsets of condition attributes of the same size.

An inspection of the $(l,\ u)$-graph for $\sigma^{l,u}(\{c_1, c_3\}, D)$ shows it contains shaded regions with $\sigma^{l,u}(C,\ D) < 1$, no other subset of condition attributes has this shaded region. Hence in Table 3 this also eludes to why $PoD_{QoS}(\{c_1,c_3\}) > PoA_{QoS}(\{c_1,c_3\})$ and not for any other subset of condition attributes.

5 Conclusion

This paper introduces the degree of sagacity measure in $\text{VPRS}_{l,u}$. One reason for this investigation is the acknowledgement in the different levels of allowed miss-classification in the possible classification of an object to a decision class or its compliment. How important the notion of the classification to the compliment of a decision class needs to be elucidated.

In the wine problem considered here with three possible decision classes, classification to the compliment of a decision class could suggest classification to either of the other two decision classes. Should we conclude this or stop at saying only it is the compliment of the decision class. This relates to the notion of open and closed world cases. That is, in the open world case there may exist other categories not included in the decision table in question - this may be a sample decision table from a population. Hence only the compliment of a decision class can be considered. The closed world case would allow the next stage that the classification is instead to one of the other decision classes included in the decision table.

Within Dempster-Shafer theory (DST) the issue of open and closed worlds is a conspicuous issue. That is, the extant literature of DST considers how these two separate cases should be approached. The question here is can RST incorporate formally the notion of open and closed world cases. In particular, $\text{VPRS}_{l,u}$ with its added dimension of miss-classification to the compliment of a decision class may be one direction to consider this problem. Indeed with an l value specific to the allowance of miss-classification to the compliment of a decision class its value could suggest whether open or closed world is being considered.

References

[1] Beynon, M.: Investigating the choice of l and u values in the extended variable precision rough sets model. Rough Sets and Current Trends in Computing RSCTC2002, Penn State University USA (2002) 61-68

[2] Fayyad, U.M., Irani, K.B.: On the handling of continuous-valued attributes in decision tree generation, Machine Learning 8 (1992), 87-102.

[3] Katzberg, J.D., Ziarko, W.: Variable precision extension of rough sets. Fundamenta Informaticae 27 (1996) 155-168

[4] Ziarko, W.: Variable precision rough sets model. Journal of Computer and System Sciences 46 (1993) 39-59

Appendix A

The discretised condition attribute values and decision attribute value for the wine data are reported in Table 4.

Table 4
Condition and decision attribute values (in categorical form)

	c_1	c_2	c_3	d_1			c_1	c_2	c_3	d_1			c_1	c_2	c_3	d_1
1	1	2	1	1		15	1	2	1	2		29	1	2	1	3
2	1	2	2	1		16	1	1	1	2		30	2	2	1	3
3	1	2	1	1		17	1	1	1	2		31	2	2	1	3
4	1	2	2	1		18	2	2	1	2		32	2	2	1	3
5	1	2	2	1		19	1	2	1	2		33	2	2	2	3
6	1	2	1	1		20	1	2	1	2		34	2	2	1	3
7	1	2	2	1		21	1	2	1	2		35	1	2	2	3
8	1	2	1	1		22	1	2	1	2		36	2	2	2	3
9	2	2	1	1		23	1	2	2	2		37	2	2	2	3
10	2	2	2	1		24	1	1	1	2		38	2	2	1	3
11	1	2	2	1		25	1	2	1	2		39	2	2	2	3
12	2	2	2	1		26	1	2	1	2		40	2	2	2	3
13	1	2	2	1		27	1	1	1	2						
14	1	2	2	1		28	2	2	2	3						