Knowledge-Based Systems in Bioinformatics

Lab 2 – Introduction to Rough Sets

Teachers:

Doğa Yalova	doga.yalova.3905@student.uu.se
Matthew Redmayne	matthew.redmayne.0786@student.uu.se
Lukas Bleichner	lukas.bleichner.5753@student.uu.se
Girish Pulinkala	girish.pulinkala@icm.uu.se

Introduction

Please submit your answers (in pdf) through your student portal account.

Consider the decision system $A = (U, A \cup \{d\})$

	Age	***Sex***	***LEMS***	***Walk***
Smith	16-30	Male	50	**Yes**
Jones	16-30	Male	0	**No**
Parker	31-45	Male	26-49	**No**
Hanson	31-45	Male	26-49	**Yes**
Moore	46-60	Female	1-25	**No**
Fields	16-30	Female	50	**Yes**
Starr	46-60	Female	1-25	**No**

Tasks

1. Formal definitions.

In the example above, what are the following: The universe U? The set of attributes A? All sets V_a?

2. Discernibility I

(a) What is the set $IND\{Age\}$?

(b) What are the *upper* and *lower approximations* of *Walk=Yes*, given only the attribute *Age*? How accurate the approximation is?

(c) What is the value of the *rough membership function* for each of the members of $IND\{Age\}$? Is this a good model of *Walk=Yes*?

3. Discernibility II

(a) Let B={Age,Sex}. What is the positive region of B, $POS_B(\{d\})$?

(b) Let A={Age,Sex,LEMS}. What is the positive region of A, $POS_A(\{d\})$?

4. Discernibility III

(a) Compute the equivalence classes of the given information system, i.e., without the decision attribute.

(b) Compute the entries of the discernibility matrix with respect to the condition attributes.

(c) Compute the Boolean discernibility function. How do you interpret the obtained Boolean function? Apply it on the decision system to find whether (Smith, Parker, and Hanson) are discernible from each other.

5. Rule-based classification

(a) For each equivalence class in 4. a, compute its value for the generalized decision attribute. Is the table consistent?

(b) Compute the decision-relative discernibility matrix.

(c) Compute the Boolean discernibility function that expresses how all decision classes can be

discerned from each other. State its prime implicants.

(d) What decision rules can be defined from the reduct(s) obtained in task (c)?

(e) Compute Accuracy, Coverage, Strength, and Support for a rule from the set of rules defined in task (d) where the decision attribute (WALK) = No.

6. Classifier Performance

Assume that we have trained a classifier and tested it, with the following results:

	Classifier says “Lung” (1)	Classifier says “Colon” (0)
Real origin “Lung” (1)	34	3
Real origin “Colon” (0)	10	22

Let’s say that we consider “Lung” as positive cases and “Colon” as negative. Then, what are the values of the following:

(a) True positives?

(b) True negatives?

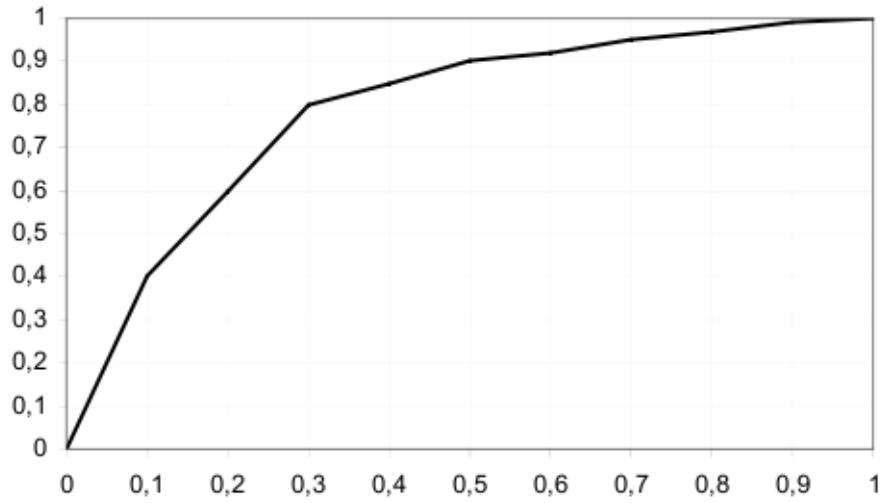
(c) False negatives?

(d) False positives?

(e) Sensitivity?

(f) Specificity?

Now consider the following ROC curve:



(g) How do you interpret this curve: What should the labels be on the X- and the Y-axis?

(h) What can you say about the performance of this classifier?

(i) What is a good threshold value? (Discuss in general how to choose a threshold)