A rough sets based classifier for primary biliary cirrhosis.

Kenneth Revett
Harrow School of Computer Science, University of Westminster

A Rough Sets Based Classifier for Primary Biliary Cirrhosis

Kenneth Revett

Abstract — **In this paper, a decision support system is presented based on the machine learning approach of rough sets. The resulting decision support system was able to reduce the dimensionality of the data, produce a highly accurate classifier, and generate a rule based classifier that is readily understood by a domain expert. These preliminary results indicate that the rough sets machine learning approach can be successfully applied to biomedical datasets that contain a variety of attribute types, missing values and multiple decision classes.**

Keywords **cirrhosis diagnosis, dimensionality reduction, medical decision support systems, and rough sets**

I. Introduction

Primary biliary cirrhosis (PBC) is a disease characterized by inflammatory destruction of the small bile ducts within the liver, which eventually leads to cirrhosis of the liver. The cause of PBC is unknown, but because of the presence of autoantibodies, it is generally thought to be an autoimmune disease [1]. Other etiologies, such as infectious agents, have not been completely excluded. PBC has a worldwide prevalence of approximately 5/100,000 and an annual incidence of approximately 6/1,000,000. The prevalence and incidence appear to be similar in different regions of the world. About 90% of patients with PBC are women. Most commonly, the disease is diagnosed in patients between the ages of 40 and 60 years. Currently, there is no cure for this disease, although D-penicillamine has been tried in clinical trials [2]. These clinical trials have led to publicly available datasets of patients with PCB. In this study, we examine a publicly available dataset on 312 patients diagnosed with PCB at various stages during disease progression. This dataset contains a series of attributes and uniquely, contains data on multiple hospital visits. What we wish to derive from this dataset are correlations between the attributes which are a series of clinically relevant variable measurements that are known to be relevant to the disease - and the decision outcome. In addition, we wish to remove any variables/attributes that appear to be non-informative for the purposes of diagnosing PCB or at least are not relevant to the clinical outcome. The approach taken in this paper is that of the machine learning paradigm of Rough sets, first proposed by Pawlak in the early 1990s as a means of extracting knowledge from data data [3]. Rough sets have been used in as a tool to investigate the relationships between attributes and clinical outcomes across a variety of biomedical datasets. [4]-[6]. One of the hallmark features of rough sets is the ability to remove redundant attribute [7]. In addition, rough sets provides a highly accurate classification system that is rule based. In this paper, we utilise these features of rough sets to data-mine the PCB dataset. This paper is organised as follows: in the next section we present a brief description of the rough set algorithm, followed by a description of the dataset, then a results section followed by a conclusion and future work.

Kenneth Revett is with the Harrow School of Computer Science, University of Westminster, London, UK (phone: +442079115000, fax: +442079115608, email: revettk@westminster.ac.uk.

II. Rough Sets

Our PCB classifier is based on the concept of approximate reducts derived from the data-mining paradigm of the theory of Rough Sets [3],[7]. The dataset consists of a number of attributes (18) and a decision for each patient. We used these datasets to generate a set of rules of the form "if (Attribute 1 = X) and (attribute 2 = Y) => decision = Z". These rules are generated automatically through the application of the rough set algorithm (we used the Rosetta implementation) [8]. We divide the decision table into training and test cases, employing N-fold cross validation. The data set is transformed into a decision table (DT) from which rules are generated to provide an automated classification capacity. In generating the decision table, each row consists of an observation (also called an object) and each column is an attribute, with the last one as the decision for this object {d}. Formally, a DT is a pair $A = (U, A \cup \{d\})$ where $d \notin A$ is the *decision attribute,* where U is a finite non-empty set of objects called the *universe* and A is a finite non-empty set of attributes such that $a:U \rightarrow V_a$ is called the value set of a. Rough sets seeks data reduction through the concept of equivalence classes (through the indiscernibility relation). By generating such classes, one can reduce the number of attributes in the decision table by selecting any member of the equivalence class as a representation of the entire class. This process generates a series of *reducts* which are subsequently used in the classification process. Finding the reducts is an NP-hard problem, but fortunately there are good heuristics that can compute a sufficient amount of reducts in reasonable time to be usable. In the software system that we employ an order based genetic algorithm (o-GA) [9] which is used to search through the decision table for approximate reducts which result in a series of 'if..then..' decision rules. We then apply these decision

1-4244-0049-X/05/$20.00 ©2005 IEEE

rules to the test data and measure specificity and sensitivity of the resulting classifications.

III. DECISION TABLE DESCRIPTION

The PCB dataset consists of 312 records of patients that were diagnosed with PCB at various stages during the disease development. A number of attributes were measured and recorded (18), as clinically relevant measurements over a period of several months to years. Table 1 below presents a listing of the attributes and their value ranges. Each patient was seen at a medical facility on numerous occasions (on average 4 visits). In order to simplify the dataset, the average values for attributes were used, resulting in a dataset with 312 objects. Most of the attribute values were continuous, which were discretised first prior to the application of rough sets. Since there were many repeat visits to medical facilities, many of the tests were not performed at all visits, and hence yield a null value in the dataset. Generally speaking, null or missing values reduce the information content of the dataset but in this case, they may also serve to indicate that a particular test was not necessary. This may have medical implications, but they are not explored in this preliminary study. Therefore, missing values were imputed when necessary using a conditioned median fill method available in Rosetta., yielding a complete dataset to work with. The decision class was multi-values, with '0' = alive,, '1' = transplanted, and '2' = dead. The details of the final decision table are listed in Table 1.

IV. METHODS

With a fully complete decision table, we proceeded to apply the rough sets algorithm on the dataset. As a first step, we discretised the data using an entropy preserving/MDL (minimal description length) algorithm. This produces a complete discretisation of all of the continuous attributes (please see Table I for details on the type of attributes included in this decision table). The same discretisation was applied to the training/test cases. Next reducts were generated using a genetic algorithm based search technique. The resulting reducts were used to generate decision rules, by mapping the attribute values directly onto the decision table and reading off the resulting classification value. In order to determine the accuracy of the classification task, the sensitivity and specificity values were measured, along with the positive predictive value (PPV) and the predictive negative value (PNV). Confusion matrices and ROC curves can be automatically generated in Rosetta (see Table III below for a set of representative confusion matrices). These results provide quantitative data that can be used to compare various machine learning algorithms in a unified and consistent fashion, allowing users to select the optimal approach for a given class of problems. These steps were repeated 10 times, randomly selecting 70% (218) entries with replacement for training and the balance (94 entries) for testing purposes. Lastly, we present a sample of rules that represent the result of the classification algorithm.

V. RESULTS

TABLE 1: PBC datasets description. Please consult [2] for more details on the dataset

Case Number	Integer (continuous)
Days since registration	Integer (continuous)
Drug	1 = D-penicillamine, 0 = Placebo
Age at initial registration	Integer
Sex	0 = male, 1 = female
Days between study enrollment and a visit	Integer (continuous)
Presence of ascites	0 = no, 1 = yes
Presence of hepatomegaly	0 = no, 1= yes
Presence of spiders	0 = no, 1 = yes
Presence of edema	0 = no, 1 = yes
Serum bilirubin	Mg/dl (continuos)
Serum cholesterol	Mg/dl (continuous)
Albumin	Gm/dl (continuous)
Alkaline phosphatase	U/liter (continuous)
SGOT	U/ml (continuous)
Platelets	Ml^3/1000
Prothrombin time	Seconds (continuos)
Histologic stages of disease	Discrete (1-4)
Classification	0 = alive, 1 = transplant, 2 = dead

TABLE 2: Pearson Correlation coefficients for all attributes used in the decision table (excluding the decision class)

Case Number	-0.26
Days since registration	-0.51
Drug	-0.01
Age at initial registration	0.18
Sex	-0.19
Days between study enrollment and a visit	-0.19
Presence of ascites	0.20
Presence of hepatomegaly	0.32
Presence of spiders	0.22
Presence of edema	0.28
Serum bilirubin	0.38
Serum cholesterol	0.17
Albumin	-0.30
Alkaline phosphatase	0.23
SGOT	0.28
Platelets	-0.14
Prothrombin time	0.24
Histologic stages of disease	0.23

The results from Table 2 indicate that there are no strongly correlated attributes with respect to the decision classes: generally a large (greater than |0.7| is indicative of a strong correlation). We were therefore not able to, by direct

inspection of the Pearson correlation coefficients, remove any attributes directly from the decision table. We proceeded to produce reducts, of which there were 37, with an average length of 3 attributes. From these reducts, Rosetta generated the decision rules that would be used to verify the accuracy on the training set and were then used for the classification of the testing

TABLE 3: Sample confusion matrices randomly selected from a series of 10 classifications. Please note all calculated values are truncated with rounding to two decimal places.

	Alive	Transplant	Dead	
Alive	46	0	0	1.0
Transplant	1	14	1	0.88
Dead	0	2	30	0.94
	0.98	0.88	0.97	**0.96**
Alive	44	2	0	0.96
Transplant	2	13	1	0.81
Dead	2	1	29	0.91
	0.92	0.81	0.97	**0.91**
Alive	45	1	0	0.98
Transplant	1	15	0	0.94
Dead	0	3	29	0.91
	0.98	0.80	1.0	**0.95**

In table 3, above, the confusion matrix bold values at the bottom right hand corners of each confusion matrix entry is the overall accuracy, according to the following formula:

$$Acc = TP + TN /(TP + FP + TN + FN) \quad (1)$$

As can be seen, the average accuracy was approximately 94% (area under the ROC = 97%). As can be seen in Table 3, the results are slightly skewed to lower values as a result of the 'transplant' group. This may reflect the fewer sample numbers for this group compared to the other two categories. If we excluded the 'Transplant' category, the classification accuracy approaches 98%. These results indicate that rough sets is able to classify the data quite accurately. The last issue to be investigated in this preliminary study is the rule set: one of the hallmarks of rough sets is its facility to generate easy to understand rules. These rules are in the form of traditional conjunctive normal for: if 'Attribute X = A' ^ 'Attribute Y = B => decision = Q. In order for a rule based classifier to be useful, the rules must be interpretable, at least for a domain expert. We evaluated the rules with respect to a complexity measure which is essentially their length and their support. In rough sets, rules of varying lengths can be generated, depending on the dataset under examination and the parameters used to generate the rule set. Generally, shorter rules are more appropriate than longer rules with respect to their generalisability capacity. Long complicated rules tend to imply that the classifier has overfit the data and tend not to generalise well to new test cases. What is generally sought are short rules, with respect to the number of attributes they contain. Analogous to the building block hypothesis of Goldberg, we want short and highly accurate rules. In Table 4 below, we present a small sample of the rules that were generated with our rough sets classifier.

Table 4. As ample of the rules produced by the rough sets classifier. The rules combine attributes in conjunctive normal form and map each to a specific decision class. The '*' corresponds to an end point in the discretised range the lowest value if it appears on the left hand side of a sub-range or the maximum value if it appears on the right hand side of a sub-range

Antecedents => Consequent
Enroll days([*, 189)) AND asictes(0) => Decision(0)
Edema(0.0) AND Platelets([218, 423)) => Decision(0)
Albumin([3.56, 3.92)) AND Platelets([218, 423)) => Decision(1)
Albumin([*, 3.11)) AND Platelets([*, 133)) => Decision(2)
Asictes(1) AND Platelets([*, 133)) => Decision(2)

Please note that decision class '0' corresponds to 'Alive,' '1' corresponds to 'Transplant,' and '2' corresponds to 'Death.' The rules from this dataset tended to have a defining length on average of 3. In addition, only 5 out of the total attribute set was obtained in the rule set, resulting in a significant simplification of the decision table (5/18). This is one of the hallmark features of the rough sets paradigm the automated dimensionality reduction of attributes from decision tables. The significance of this result is that the dataset appears to contain many attributes that are not related to the decision class. This result means we can reduce this dataset considerably, without losing valuable information. In addition to the complexity of the rules, another issue is the number of rules that are generated.. if the rules are going to form the basis for a rule-based expert system, it would be useful if the number of rules was kept to a minimum in order to reduce the computational expense of searching through the rule base in order to answer a query. A useful feature of Rosetta is the ability to filter rules on several criteria: right hand side (RHS) support, RHS coverage, rules with specific decisions, rules with LHS length within a given range. Support is a measure of the number of objects in the decision table that match a certain rule (that is their antecendent and consequent(s) are matched within the decision table). It is a useful measure of the relative importance of a given object within the decision table. In Table 5, data is presented relating the quality of the classification with respect to the number of rules. The rules were filtered based on RHS support (specified as a range) and mapped against the resulting classification accuracy. What was sought was a reduction in the cardinality of the antecedents without significantly reducing the classification accuracy. Through empircal exploration, it was found that specifying a RHS support bewteen 2-4 reduced the cardinality of the rules significantly without a concomitant loss in classification accuracy. These results are depicted in Table 5 below.

Table 5: Sub-range of RHS support, number of rules, and the resulting accuracy

RHS Support	Number of Rules	Accuracy
0	5702	95%
0-2	629	93%
2-4	314	83%
4-6	189	77%

IV. CONCLUSION

In this study, we present a preliminary study of an internet housed dataset containing 312 records of patients that have been diagnosed with primary biliary cirrhosis. The dataset required a considerable amount of pre-processing as it contained multiple visits for each patient, which meant that some tests were not always performed for each visit. At this stage of the investigation, the multiple visits were handled by averaging the values that were obtained if there were multiple values for the same attribute (i.e. test measured on more than one occasion). Even with this simplifying assumption, rough sets was able to generate a highly accurate classifier, compared with other results [10],[11]. An accuracy of 95% was quite positive (97% area under the ROC curve). Rough sets was also able to reduce the attributes to five (ascites, edema, platelets, albumin, and enroll days). Although the data was not presented, when all but these five attributes were masked from the decision table and the entire rule generation process completed as in the control case, there was not significant change in the classification accuracy. Unfortunately, none of the attributes appeared to be highly correlated with the decision class (as per Table 2). Rough sets generates a set of easy to interpret rules that can be directly useful to a person with the appropriate domain knowledge. These rules relate directly to the attributes and can map to the appropriate decision class. In this decision table, there are three different decision outcomes, in principle there can be as many as one wishes there are no theoretical limits here. What is important in rough sets is the number of objects of a given decision class in this particular dataset, there were only 32 objects with decision class '1' that is 'transplant.' In this case, the resulting classification accuracy tends to be reduced when compared to cases where the number of objects for a class is large compared with the number of attributes. Currently, there is no specific ratio based on theoretical principles this is an area of active investigation when dealing specifically with rough sets. Lastly, the number of rules generated although fairly high is still manageable by today's computational capacities. The rule set can be easily integrated into an expert system forming the basis of a powerful medical expert diagnosis facility. This area will be explored in further efforts on this dataset.

REFERENCES

[1]

[2] Tromm, B. May, C.G. Klein,A. Fissler-Eckhoff, & T. Griga.., "Long-term response of primary biliary cirrhosis (stage I) to therapy with ursodeoxycholic acid." Hepatogastroenterology. May-Jun;52(63):753-6, 2005.

[3] Howard J. Worman, M. D: http://cpmcnet.columbia.edu/dept/gi/PBC.html

[4] Z. Pawlak . Rough Sets, International Journal of Computer and Information Sciences, 11, pp. 341-356, 1982.

[5] Revett, K. & Kahn, A. A rough Sets Based breast Cancer Decision Support System, METMBS, Las Vegas, Nevada, USA, June 17-19, 2005

[6] Khan & K. Revett. "Data mining the PIMA Indian diabetes database using Rough Set theory with a special emphasis on rule reduction," INMIC2004, Lahore Pakistan, pp. 334-339, December, 2004.

[7] Revett, K. & Kahn, A. Data-mining Small Biomedical Datasets Using Rough Sets, MCHC2005 Conference, Craiova, Romania, pp. 231-241, 2005.

[8] Slezak, D. .Approximate Entropy Reducts, Fundamenta Informaticae, 2002.

[9] Rosetta: Rosetta: http://www.idi.ntnu.no/~aleks/rosetta

[10] Wroblewski., J. Theoretical Foundations of Order-Based Genetic Algorithms, Fundamenta Informaticae 28(3-4) pp. 423–430, 1996.

[11] Murtaugh PA. Dickson ER. Van Dam GM. Malinchoc M. Grambsch PM. Langworthy AL. Gips CH. Primary biliary cirrhosis: prediction of short-term survival based on repeated patient visits, Hepatology. 20(1.1):126-34, 1994.

[12] Markus, et al., N Eng J of Med 320:1709-13, 1989.