Data Mining a Prostate Cancer Dataset Using Rough Sets

Kenneth Revett , Sérgio Tenreiro de Magalhães, and Henrique M. D. Santos, member IEEE

Abstract— **Prostate cancer remains one of the leading causes of cancer death worldwide, with a reported incidence rate of 650,000 cases per annum worldwide. The causal factors of prostate cancer still remain to be determined. In this paper, we investigate a medical dataset containing clinical information on 502 prostate cancer patients using the machine learning technique of rough sets. Our preliminary results yield a classification accuracy of 90%, with high sensitivity and specificity (both at approximately 91%). Our results yield a predictive positive value (PPN) of 81% and a predictive negative value (PNV) of 95%. In addition to the high classification accuracy of our system, the rough set approach also provides a rule-based inference mechanism for information extraction that is suitable for integration into a rule-based system. The generated rules relate directly to the attributes and their values and provide a direct mapping between them.**

Index Terms— cancer classifier ,machine learning, prostate cancer dataset, reducts, Rough sets

I. Introduction

Prostate cancer is the second leading cause of mortality in men, exceeded only by lung cancer. The cause(s) of this form of cancer remain to be elucidated, but factors such as diet, heredity, and environmental factors that effect male hormones (androgens) have been implicated in epidemiological studies [1,3,5]. Currently, two standard tests are used for early detection of prostate cancer:

- Digital rectal examination (DRE). With the DRE, a physician palpates the prostate in order to feel lumps or masses.
- PSA test. The PSA blood test measures the level of a protein called prostate-specific antigen. It is able to detect early prostate cancer, although it has limitations.

There are many unresolved questions surrounding PSA testing. The test is not accurate enough to either completely rule out or confirm the presence of cancer. Current treatments entail chemotherapy, surgery or a combination of the two depending on the stage of disease progression. Further, the incidence of prostate cancer increases with age. This will present an increased incidence as the world population tends towards increased longevity. This trend is alarming and warrants investigating the causative factors in prostate cancer through all available means. In this paper, we present the results of a machine learning technique based on rough sets to the study of a clinically relevant prostate cancer dataset.

In this study, we investigate a dataset containing data on 502 (29%/71% live/dead) patients that were diagnosed with prostate cancer. The dataset contains 18 attributes including the decision attribute (see section 2.1 for a listing of the attributes) with 27 missing values (0.3%). We investigated this dataset with respect to the following: i) attribute pruning, ii) classification accuracy and iii) rule induction. Pruning (dimensionality reduction) removes variables that are not directly related to the classification process. This feature of rough sets makes the dataset much easier to work with and may help to highlight the relevant classification features of the data. Once the redundant features have been pruned from the dataset, rough sets is used in the classification process, mapping attributes and their values to decision classes. In many cases, rough sets are able to produce classification accuracy that is superior to more 'traditional' classification algorithms. Lastly, rough sets provide a set of decision rules that are readily interpretable by a domain expert. These rules map attributes and their values to decision classes. These three facilities available in the rough set paradigm provide a unique and consistent approach to extracting knowledge from data. In the next section, we present an overview of rough sets, followed by the use of rough sets to classify this particular dataset, followed by a results section and lastly a summary of this work.

rostate cancer is the second leading cause of mortality in men, exceeded only by lung cancer. The cause(s) of this form of cancer remain to be elucidated, but factors such as diet, heredity, and environmental factors that effect male hormones (androgens) have been implicated in epidemiological studies [1,3,5]. Currently, two standard tests are used for early detection of prostate cancer:

of cancer. Current treatments entail chemotherapy, surgery or a combination of the two depending on the stage of disease progression. Further, the incidence of prostate cancer increases with age. This will present an increased incidence as the world population tends towards increased longevity. This trend is alarming and warrants investigating the causative factors in prostate cancer through all available means. In this paper, we present the results of a machine learning technique based on rough sets to the study of a clinically relevant prostate cancer dataset.

K Revett is with the Harrow School of Computer Science, University of Westminster, London, UK (phone: 44-2079115000; fax: 44-2079115609; e-mail: revettk@westminster.ac.uk).

S. Tenreiro de Magalhães Gorunescu, is with the Department of Information Systems, University do Minho, Guimaraes, Portugal (e-mail: psmagalhaes@dsi.uminho.pt);

H. MD Santos, is with the Department of Information Systems, University do Minho, Guimaraes, Portugal (e-mail: hsantos@dsi.uminho.pt); *IEEE member*

II. Introduction to Rough Sets

Rough set theory is a relatively new data-mining technique used in the discovery of patterns within data first formally introduced by Pawlak in 1982 [10,11]. Since its inception, the rough sets approach has been successfully applied to deal with vague or imprecise concepts, extract knowledge from data, and to reason about knowledge derived from the data [12,14]. We demonstrate that rough sets has the capacity to evaluate the importance (information content) of attributes, discovers patterns within data, eliminates redundant attributes, and yields the minimum subset of attributes for the purpose of knowledge extraction.

The first step in the process of mining any dataset using rough sets is to transform the data into a decision table. In a decision table (DT), each row consists of an observation (also called an object) and each column is an attribute, one of which is the decision attribute for the observation {d}. Formally, a DT is a pair A = (*U*, A∪{d}) where d ∉ A is the *decision attribute*, *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->V_a is called the value set of a. Once the DT has been produced, the next stage entails cleansing the data.

There are several issues involved in small datasets – such as missing values, various types of data (categorical, nominal and interval) and multiple decision classes. Each of these potential problems must be addressed in order to maximise the information gain from a DT. Missing values is very often a problem in biomedical datasets and can arise in two different ways. It may be that an omission of a value for one or more subject was intentional – there was no reason to collect that measurement for this particular subject (i.e. 'not applicable' as opposed to 'not recorded'). In the second case, data was not available for a particular subject and therefore was omitted from the table. We have 2 options available to us: remove the incomplete records from the DT or try to estimate what the missing value(s) should be. The first method is obviously the simplest, but we may not be able to afford removing records if the DT is small to begin with. So we must derive some method for filling in missing data without biasing the DT. In many cases, an expert with the appropriate domain knowledge may provide assistance in determining what the missing value should be – or else is able to provide feedback on the estimation generated by the data collector. In this study, we employ a conditioned mean/mode fill method for data imputation. In each case, the mean or mode is used (in the event of a tie in the mode version, a random selection is used) to fill in the missing values, based on the particular attribute in question, conditioned on the particular decision class the attribute belongs to. There are many variations on this them, and the interested reader is directed to ([8,9,13]) for an extended discussion on this critical issue. Once missing values are handled, the next step is to discretise the dataset.

The basic philosophy of rough sets is to reduce the elements (attributes) in a DT based on the information content of each attribute or collection of attributes (objects) such that the there is a mapping between similar objects and a corresponding decision class. In general, not all of the information contained in a DT is required: many of the attributes may be redundant in the sense that they do not directly influence which decision class a particular object belongs to. One of the primary goals of rough sets is to eliminate attributes that are redundant. Rough sets use the notion of the lower and upper approximation of sets in order to generate decision boundaries that are employed to classify objects. Consider a decision table A = (U, A∪{d}) and let $B \subseteq A$ and X ⊆ U. What we wish to do is to approximate X by the information contained in B by constructing the B-lower (B_L) and B-upper (B^U) approximation of X. The objects in B_L (B_LX) can be classified with certainty as members of X, while objects in B^U are not guaranteed to be members of X. The difference between the 2 approximations: B^U - B_L, determines whether the set is rough or not: if it is empty, the set is crisp otherwise it is a *rough set*. What we wish to do then is to partition the objects in the DT such that objects that are similar to one another (by virtue of their attribute values) are treated as a single entity. One potential difficulty arises in this regard is if the DT contains inconsistent data. In this case, antecedents with the same values map to different decision outcomes (or the same decision class maps to two or more sets of antecedents). There are means of handling this and the interested reader should consult [2,4] for a detailed discussion of this interesting topic. The next step is to reduce the DT to a collection of attributes/values that maximises the information content of the decision table. This step is accomplished through the use of the indiscernibility relation *IND(B)* and is defined for any subset $B \subseteq A$ ($B \subseteq A \cup \{d\}$) as follows:

$$\textbf{\textit{IND}(\textit{B})=} \left\{ (x,y) \in U \times U \text{:for every } a \in B \; a(x) = a(y) \right\} \quad \textbf{(1)}$$

The elements of *IND(B)* correspond to the notion of an equivalence class. The advantage of this process is that any member of the equivalence class can be used to represent the entire class – thereby reducing the dimensionality of the objects in the DT. This leads directly into the concept of a *reduct*, which is the minimal set of attributes from a DT that preserves the equivalence relation between conditioned attributes and decision values. It is the minimal amount of information required to distinguish objects with in U. The collection of all reducts that together provide classification of all objects in the DT is called the *CORE*(A). The CORE specifies the minimal set of elements/values in the DT which are required to correctly classify objects in the DT. Removing any element from this set reduces the classification accuracy. It should be noted that searching for minimal 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) which is used to search through the decision table for approximate reducts [15]. The reducts are approximate because we do not perform an exhaustive search via the o-GA which may miss one or more attributes that should be included as a reduct. Once we have our set of reducts, we are ready to produce a set of rules that will form the basis for object classification.

Rough sets generates a collection of 'if..then..' decision rules that are used to classify the objects in the DT.

These rules are generated from the application of reducts to the decision table, looking for instances where the conditionals match those contained in the set of reducts and reading off the values from the DT. If the data is consistent, then all objects with the same conditional values as those found in a particular reduct will always map to the same decision value. In many cases though, the DT is not consistent, and instead we must contend with some amount of indeterminism. In this case, a decision has to be made regarding which decision class should be used when there are more than 1 matching conditioned attribute values. Simple voting may work in many cases, where votes are cast in proportion to the support of the particular class of objects. In addition to inconsistencies within the data, the primary challenge in inducing rules from decision tables is in the determination of which attributes should be included in the conditional part of the rule. If the rules are too detailed (i.e. they incorporate reducts that are maximal in length), they will tend to overfit the training set and classify weakly on test cases. What are generally sought in this regard are rules that possess low cardinality, as this makes the rules more generally applicable. This idea is analogous to the building block hypothesis used in genetics algorithms, where we wish to select for highly accurate and low defining length gene segments [14]. Discussion of these ideas is beyond the scope of this paper and the interested reader is directed towards [12] for a detailed discussion of these alternatives.

The rules that are generated are in the traditional conjunctive normal form and are easily applied to the objects in the DT. What we are interested in is the accuracy of the classification process – how well has the training rule set classified new objects? In addition, what sort of confidence do we have in the resulting classification of particular validation training set?

Table 1. The dataset used in this study. The numbers in parentheses refer to the number of categories for that particular attribute. For details on category values, see reference [3,6].

Attribute Name	Attribute Type
Patient number	Double
Stage	Double
Treatment	Integer (4)
Dtime(follow up time in months)	Double
Date on Study	Double
Age	Integer
Weight index (wt(kg) – ht(cm) + 200)	Integer
Pf	Integer (4)
Hx (history of cardio-vascular disease)	Double
Sbp Systolic bp	Double
Dbp Diastolic bp	Double
EKG	Integer (7)
Hg (serum haemoglobin (g/100ml)	Double
Sz (size of primary tumour (cm^2)	Double
Index of Stage and Histology	Double
Serum Prostatic Acid Phosphatase	Double
Bone Metatastases	Double
Status	Double (10)

These are standard issues that hold true for any machine learning application. In addition questions arise regarding methods for handling biomedical datasets that contain an unequal distribution of decision class objects. Traditionally in rough sets, validation is accomplished through N-fold validation, where the N is dependent upon the particular dataset at hand – but generally a 70/30 training/validation scheme is used, with replication with replacement on the order of 10% of the sample size.

In the next section, we describe the dataset that was used in this study. In addition, we describe how we analysed this dataset with respect to handling missing values, discretisation and our validation procedure strategy.

III METHODS

The structure of the dataset consisted of 18 attributes, including the decision attribute (labelled 'status') which is displayed for convenience in table 1 above. There were 502 entries in the table with only 27 missing values. We employed a 'filling in' technique to complete the dataset by using a conditioned mean fill algorithm. Essentially, each missing attribute is replaced by the mean for those attributes that belonged to the same decision class. Once all of the missing values were replaced, we then proceeded to discretise the dataset. Since rough sets works ideally with categorical data, we discretised all ordinal attributes in order to generate a completely categorical dataset. We discretised the following attributes: sbp, dbp, Hg, Sz, and Index of Stage and Histology attributes into 3 bins using equal frequency binning. We also removed by masking the following attributes: Patient number and date on Study. We also determine the Pearson's Correlation Coefficient of each attribute with respect to the decision class. The correlation values can be used to determine if one or more attributes are strongly correlated with a decision class. In many cases, this feature can be used to reduce the dimensionality of the dataset prior to classification. With these pre-processing steps completed, we then applied the rough sets algorithm to the dataset. After several experiments, we decided to use dynamic reducts based on the resulting classification accuracy. With the collection of dynamic reducts, we went on to produce the final classification. We tried variations in the number of training and testing objects, and found that a 70/30 split provided the best result. We repeated the classification process 20 times, selecting randomly with replacement. The results we obtained are described in the next section.

IV RESULTS

The classification accuracy obtained in this study was significantly affected by the extent of the pre-processing procedure. Without any pre-processing at all, we obtained an average classification accuracy of approximately 60% (10 trials). We therefore pre-processed the dataset according to the strategies specified in the methods section described above. We first calculated the Pearson Correlation coefficients for all attributes in the decision table (excluding the decision attribute). We did not find any attributes that were strongly correlated with a particular decision class, although the Dtime attribute yielded a large negative correlation (-0.74). The summary results for the correlation analysis are displayed in table 2 below. From our experience, attributes with very low correlation coefficients (positive or negative) can be removed from the decision table without compromising classification accuracy [7,12]. This particular dataset yielded very low correlation coefficients for virtually all attributes, indicating that the decision class was not heavily weighted towards any particular attribute(s). We therefore did not exclude any attributes when generating decision rules based on the value of the correlation coefficient.

We generated dynamics reducts from the DT – as this method generally performs best when the data has a reasonably large number of attributes (8,534 in this dataset). We then generated rule from the dynamic reducts, followed by classification of the test cases (25% - 125 objects were randomly selected with replacement for classification purposes). Table 3 presents five randomly selected confusion matrices that we generate from the classification procedure. The overall average classification accuracy was 89.6%, which is considerably better than other reported results on this dataset of 80 and 83% respectively [6,14].

The classification accuracy may have biased by the low number of 'live' patient cases in this study (29% or 146/502 cases) – resulting in low false positives (last entry in each test result under the 'Alive' column heading in Table 3).

Table 3. **Confusion matrices from a set of five randomly selected classification tasks on the test case (using 70/30 train/test) from 20 randomly selected tests**

Test1	Alive	Dead	
Alive	34	3	0.918919
Dead	7	81	0.920455
	0.829268	0.964286	0.92
Test2			
Alive	28	3	0.903226
Dead	11	83	0.882979
	0.717949	0.965116	0.888
Test3			
Alive	39	7	0.847826
Dead	9	70	0.886076
	0.8125	0.909091	0.872
Test4			
Alive	37	4	0.902439
Dead	10	74	0.880952
	0.787234	0.948718	0.888
Test5			
Alive	35	4	0.897436
Dead	15	71	0.825581
	0.7	0.946667	0.848

V. Conclusion

In this study, we examined the information content of a clinical dataset contained 18 attributes (decision class inclusive) on 502 instances of prostate cancer. We were primarily interested in how well the attributes mapped onto the decision classes. Our primary result was a classification accuracy of approximately 90% - which compares very favourably with other published reports. The results generated are in the form of a series of 'if attrA = X then decision = Y' decision rules which are readily interpretable by a domain expert. The quality of the rules generated is dependent of course on the information content of the data. In the hands of a trained person can the rules be verified in context – a task that is beyond the scope of most machine learning techniques.

III. References

[1] D.F. Andrews and A.M. Herzberg, Data, a Collection of Problems from Many Fields for the Student and Research Worker. Springer-Verlag, New York, 1985.

[2] J.G> Bazan, A. Skowron, & P.Synak,. Dynamic reducts as a Tool for Extracting Laws from Decision tables, Proceeding of the Third International Workshop on Rough Sets and Soft Computing, San Jose, California, pp 526-533, 1994.

[3] DP, Green SB: *Bulletin Cancer*, Paris 67:477-488, 1980.

[4] J. Grzyma la-Busse., "Applications of the rule induction systems LERS," In: 127, pp. 366–375, 1998.

[5] A.Jemal, T. Murray, E. Ward E, A. Samuels, Cancer Statistics, 2005. CA Cancer J Clin 55:10-30, 2005.

[6] P. Kalra, J. Togami, G. Bansal, A.W.Partin , M.K. Brawer, R.J. Babaian, L.S. Ross, & C.S. Niederberger: A Neurocomputational Model for Prostate Carcinoma Detection: Cancer, Volume 98, Issue 9 ,November 2003.

[7] 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

[8] H.S. Nguyen & A. Skowron, Quantization of real-valued attributes, proc Second International Conference on Information Science, pp 34-37, 1995.

[9] A.Øhrn,. "Discernibility and Rough Sets in Medicine" Tools and Applications. Department of Computer and Information Science. Trondheim, Norway, Norwegian University of Science and Technology: 239, 1999.

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

[11] Z. Pawlak,: Rough sets – Theoretical aspects of reasoning about data. Kluwer Publishers, 1991.L. Polkowski, L. & A. Skowron,, eds., Rough Sets in Knowledge Discovery 1: Methodology and Applications, volume 18 of Studies in Fuzziness and Soft Computing. Physica-Verlag, 1998.

[12] K. Revett. & A. Khan, Rough Sets Based Cancer Classification System, IADIS 2005

[13] M. Schemper & G. Heinze Probability imputation revisited for prognostic factor studies, *Statistics in Medicine* **16**, 73-80, 1997.

[14] A. Skowron, "Synthesis of adaptive decision systems from experimantal data (invited talk)." In: Proc.of the Fifth Scandinavian Conference on Artificial Intelligence SCAI-95 (A. Aamodt and J. Komorowski, (eds.), IOS Press Ohmsa, Amsterdam, pp.220-238., 1995.

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