Attribute extraction and classification using rough sets on a lymphoma dataset.

Kenneth Revett
Harrow School of Computer Science, University of Westminster

Nizamettin Aydin
Department of Computer Engineering, Bahcesehir University, Istanbul, Turkey

This is an electronic version of a paper presented at the International Symposium on Health Informatics and Bioinformatics: HIBIT'05, 10-12 Nov 2005, Antalya, Turkey.

Attribute Extraction and Classification Using Rough Sets on a Lymphoma Dataset

[1]*Kenneth Revett and* [2]*Nizamettin Aydin*
[1]*University of Westminster*
Harrow School of Computer Science
London, UK
revettk@westminster.ac.uk

[2]*Bahcesehir University*
Faculty of Engineering
Department of Computer Engineering
Bahcesehir 34538
Istanbul,Turkey
naydin@bahcesehir.edu.tr

Abstract

In this paper, we describe a rough sets approach to classification and attribute extraction of a small biomedical dataset. The dataset contains 148 entries with 19 attributes on patients that were suspected to have a lymphoma. Our primary goal was to be able to create a set of rules that allow the prediction of the decision class based on the values of relevant attributes. Our preliminary study of this dataset indicated that seven of the 19 attributes were predictive in this dataset. Our classification accuracy was approximately 85%, with a high sensitivity and specificity. In addition to the promising classification results, rough sets provided a means of dimensionality reduction and rule generation.

Keywords: Rough sets, lymphoma dataset, attribute extraction, dimensionality reduct, rule generation

1. Introduction

Lymphoma is a general term for a group of cancers that originate in the lymphatic system. The lymphomas are divided into two major categories: Hodgkin lymphoma and all other lymphomas, called non-Hodgkin lymphomas. Hodgkin lymphoma was named for Thomas Hodgkin, an English physician who described several cases of the disease in 1832. Hodgkin lymphoma will represent about 12.7 percent of all lymphomas diagnosed in 2004 [1]. About 62,250 Americans will be diagnosed with lymphoma in 2004. This figure includes approximately 7,880 new cases of Hodgkin lymphoma (4,330 males and 3,550 females), and 54,370 new cases of non-Hodgkin lymphoma (28,850 males and 25,520 females). The annual incidence of lymphoma has nearly doubled over the last 35 years. The cause of Hodgkin lymphoma is uncertain [1]. Many studies of environmental, especially occupational, linkages have been conducted with ambiguous results. For example, woodworking exposure has been associated with the disease, but causality has not been established. The Epstein-Barr virus has been associated with about one-third of cases of the disease. It has not been established conclusively as a cause of Hodgkin lymphoma, however. Persons infected with HTLV and HIV also have an increased probability of developing Hodgkin lymphoma [1].

In this study, we investigate a dataset containing data on 148 with four decision classes (2 normal, 81 metastases, 61 with malignant lymphoma, and 4 with fibrosis) patients that were hospitalised for suspected lymphoma. The dataset contains 19 attributes including the decision attribute (see section 2.1 for a listing of the attributes) with 0 missing values. 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.

1.1 Data mining

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 [7,8]. 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 [5,6]. 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 theme, and the interested reader is directed to [3,4] for an extended discussion on this critical issue. Once missing values are handled, the next step is to discretise the dataset. Rarely is the data contained within a DT all of ordinal type – they generally are composed of a mixture of ordinal and interval data. Discretisation refers to partitioning attributes into intervals – tantamount to searching for "cuts" in a decision tree. All values that lie within a given range are mapped onto the same value, transforming interval into categorical data. As an example of a discretisation technique, one can apply equal frequency binning, where a number of bins *n* is selected and after examining the histogram of each attribute, n-1 cuts are generated so that there is approximately the same number of items in each bin. See the discussion in [4,9] for details on this and other methods of discretisation that have been successfully applied in rough sets. Now that the DT has been pre-processed, the rough sets algorithm can be applied to the DT for the purposes of supervised classification.

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 and $X \subseteq U$. 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 ($B_L X$) 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). This is unfortunately the norm in the case of small biomedical datasets, such as the one used in this study. There are means of handling this and the interested reader should consult [6,10] 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 ( ) as follows:

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

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 [9]. 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 [11]. There are many variations on rule generation, which are implemented through the formation of alternative types of reducts such as *dynamic* and *approximate* reducts. Discussion of these ideas is beyond the scope of this paper and the interested reader is directed towards [10] 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? 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, repeated 10-20 times and the average of these runs are computed and reported. 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.

2. Methods

The structure of the dataset consisted of 19 attributes, including the decision attribute (labelled 'result') which is displayed for convenience in table 1 below. There were 2,664 entries in the table with 0 missing values. Since the data was essentially completely ordinal, no discretisation was performed on this dataset. We determined 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. We selected the attributes with the largest correlation coefficient, which left us with a total of seven attributes. In addition, we removed the 'normal' class, which only had 2 objects, leaving us with 3 decision classes. 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 also filtered the dataset based on the attributes with the highest correlation coefficients – to see if we could reduce the dimensionality of the dataset without compromising classification accuracy. We

3. Results

then repeated the classification process 20 times, selecting randomly with replacement. The results we obtained are described in the next section.

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). As a first pre-processing step, we calculated the Pearson Correlation coefficients for all attributes in the decision table (excluding the decision attribute). The summary results for the correlation analysis are displayed in Table 1 below. From our experience, attributes with very low correlation coefficients (positive or negative) can be removed from the decision table without compromising classification accuracy [4,9].

Table 1. Attributes along with the Pearson Correlation coefficient for the elements in the decision table. Please note the '*' next to the correlation values were those that were maintained in the reduced dataset

Attribute Name	**Correlation coefficient**
Lymphatics	0.0176
Block of efferents d	0.093
Block of lymphatics c	0.176 *
Block of lymphatics s	0.169 *
By pass	0.157 *
Extravasates	0.197 *
Regeneration of nodes	-0.083
Early uptake	0.251 *
Lymph node diminishing	-0.031
Lymph nodes enlarging	-0.112
Change in lymph Defect in nodes	0.101 * 0.077
Changes in nodes	-0.181
Changes in structure	-0.047
Special forms	-0.012
Dislocation	-0.060
Exclusion of lymph	0.101 *
Number of nodes	0.081

Table 2. Confusion matrices from a randomly selected set of classification tasks. We used dynamic reducts with no reduct/rule filtering, a 70/30 split on training /testing. The sensitivity is Listed in the top right column and the specificity is directly underneath the sensitivity (both are highlighted. Please note – '0' corresponds to metastases, '1' to malignant and '2' to fibrosis

Test1	**0**	**1**	**2**	
0	19	2	1	**0.863**
1	3	13	1	**0.765**
2	0	0	5	1.0
	0.863	**0.867**	**0.714**	0.845
Test2	**0**	**1**	**2**	
0	**18**	**2**	**2**	**0.818**
1	**2**	**12**	**3**	**0.706**
2	**1**	**0**	**4**	0.80
	0.818	**0.857**	**0.444**	0.7405
Test1	**0**	**1**	**2**	
0	18	3	1	**0.818**
1	3	12	2	**0.706**
2	0	1	4	0.80
	0.818	0.857	0.571	0.7616

4. Conclusion

We were able to achieve a high classification rate for this dataset, with an average of 89%. Our results provide reasonable classification accuracy, surpassing several reported values [2,3]. In the process of classifying the data, we were also able to reduce the dimensionality of the dataset to 7 attributes. In addition, rough sets generates a set of easy to interpret decision rules in the form if 'if ATTR 1 = 'X' and ATTR 2 = 'Y' then decision = 'Z'. These rules are directly interpretable by a domain expert and can serve as the basis for a decision support system. Lastly, rough sets are able to work with datasets that are small and incomplete. These properties of rough sets – makes this a very suitable tool for mining small biomedical datasets.

Acknowledgement:

The dataset was donated by**:** Igor Kononenko, University E.Kardelj, Faculty for electrical engineering, Trzaska 25 61000 Ljubljana

We also acknowledge use of the software Package Rosetta, available from the internet at: http://www.idi.ntnu.no/aleks/rosetta/

References

[1]http://www.leukemia-lymphoma.org

[2] Cestnik,G., Konenenko,I, & Bratko,I. (1987). Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.

[3] Clark,P. & Niblett,T. (1987). Induction in Noisy Domains. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 11-30,Sigma Press

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

[5] Komorowski, J., Pawlak, Z., Polkowski, L., Skowron, A.: Rough sets: A tutorial. In: S.K. Pal, A. Skowron (eds): Rough Fuzzy Hybridization – A New Trend in Decision Making. Springer Verlag (1999) pp. 3–98.

[6] 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.

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

[8] Pawlak, Z.: Rough sets – Theoretical aspects of reasoning about data. Kluwer (1991).

[9] K. Revett. & A. Khan, Rough Sets Based Cancer Classification System, IADIS 2005 (in press)

[10]. Slezak, D.: Approximate Entropy Reducts. Fundamenta Informaticae (2002).

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