Using Genetic Algorithms to Optimise Rough Set Partition Sizes for HIV Data Analysis

Bodie Crossingham and Tshilidzi Marwala

School of Electrical and Information Engineering, University of the Witwatersrand
bodie@lutrin.co.za

Summary. In this paper, we present a method to optimise rough set partition sizes, to which rule extraction is performed on HIV (Human Immunodeficiency Virus) data. The genetic algorithm optimisation technique is used to determine the partition sizes of a rough set in order to maximise the rough sets prediction accuracy. The proposed method is tested on a set of six demographic properties of individuals obtained from the South African antenatal survey, with the outcome or decision being either HIV positive or negative. Rough set theory is chosen based on the fact that it is easy to interpret the extracted rules. The prediction accuracy of equal width bin partitioning is 69.8% while the accuracy achieved after optimising the partitions is 87.5%.

1 Introduction

In the last 20 years, over 60 million people have been infected with HIV (Human Immunodeficiency Virus), and of those cases, 95% are in developing countries [1]. During this year, AIDS (Acquired Immune Deficiency Syndrome) claimed an estimated 2.9 million lives [2]. HIV has been identified as the cause of AIDS. It is thus evident that the analysis of HIV is of the utmost importance. By correctly forecasting HIV, the causal interpretations of a patients being seropositive (infected by HIV) is made much easier. Poundstone *et al* related demographic properties to the spread of HIV [3]. In their work they justified the use of demographic properties to create a model to predict HIV from a given database, as is done in this study. RST (rough set theory) uses the social and demographic factors to predict HIV status, this in turn provides insight into which variables are most sensitive in determining HIV status. Rough sets have been used in many applications. Rowland *et al* compared the use of RST and neural networks for the prediction of ambulation spinal cord injury [4], and although the neural network method produced more accurate results, its "black box" nature makes it impractical for the use of rule extraction problems. RST compromises accuracy over rule interpretability but for HIV it can be argued that interpretability of the data

is of more importance than just prediction. In order to achieve the best accuracy, the rough set partitions or discretisation process needs to be optimised. The optimisation process is done using genetic algorithm (GA), where the fitness function aims to achieve the highest accuracy produced by the rough set. Literature reviews have shown that limited work has been done on the optimisation of rough set partition sizes.

In the following sections, the background of the topic is stated, followed by a summarised explanation of RST. Section 4 explains how the genetic algorithm is used to optimise the rough set partitions, and in section 5 the results are given and compared.

2 Background

Rough set theory was introduced by Zdzislaw Pawlak in the early 1980s [5]. RST is a mathematical tool which deals with vagueness and uncertainty. Rough sets are useful in the analysis of decisions in which there are inconsistencies. To cope with these inconsistencies, lower and upper approximations of decision classes are defined [6]. One of the advantages of RST is that it does not require *a priori* knowledge about the data set, and it is for this reason that statistical methods are not sufficient for determining the relationship between the demographic variables and their respective outcomes.

The data set used in this paper was obtained from the South African antenatal sero-prevalence survey of 2001. The data was obtained through questionnaires completed by pregnant women attending selected public clinics and was conducted concurrently across all nine provinces in South Africa [7].

The six demographic variables considered are: *race, age of mother, education, gravidity, parity* and, *age of father*, with the outcome or decision being either HIV positive or negative.

The HIV status is the decision represented in binary form as either a 0 or 1, with a 0 representing HIV negative and a 1 representing HIV positive. The input data is discretised into four partitions and represented numerically as either 1, 2, 3 or 4. This number is chosen as it gives a good balance between computational efficiency and accuracy.

3 Rough Set Theory and Rough Set Formulation

Rough set theory deals with the approximation of sets that are difficult to describe with the available information [8]. Some concepts that are fundamental to RST are mentioned briefly but for a complete explanation refer to [5]. The data is represented using rows and columns in an information table. Once the table is obtained, the data is discretised into four partitions as mentioned earlier. Each case is evaluated and cases that are conflicting are referred to as an indiscernibility relation and this is the main concept of rough set theory,

(indiscernibility meaning indistinguishable from one another). RST offers a tool to deal with indiscernibility and the way in which it works is, for each concept/decision X, the greatest definable set containing X and the least definable set containing X are computed. These two sets are called the lower and upper approximation respectively.

It is through these lower and upper approximations that any rough set is defined. It must be noted that for most cases in RST, reducts are generated to enable us to discard functionally redundant information [5]. And although reducts are one of the main advantages of RST, they are ignored for the purpose of this paper, i.e. the optimisation of discretised partitions.

A membership function then determines the plausibility of which any object x belongs a particular set X. The results are illustrated using a receiver operating characteristic (ROC) curve. The ROC curve displays the relationship between sensitivity (true-positive rate) and 1-specificity (false-positive rate) across all possible threshold values that define the positivity of being infected with HIV. The area under curve (AUC) of the respective ROC curves is used as the performance criterion. The higher the AUC, the better the classification accuracy is. The AUC is the accuracy used to optimise the rough set partition sizes. The process of modelling rough sets can be broken down into five stages and can be summarised as follows:

Stage one would be to select the data. Stage two to pre-process and discretise the data. If reducts are considered, the third stage would be to use the cleaned data to generate reducts. A reduct is the most concise way in which we can discern object classes [9]. To cope with inconsistencies, lower and upper approximations of decision classes are defined [5, 6, 9]. Stage four would be to extract the generated rules and these rules can be presented in an *if* CONDITION(S)-*then* DECISION format, e.g.

- **If** Race = African **and** Mothers Age = 23 **and** Education = 4 **and** Gravidity = 2 **and** Parity = 1 **and** Fathers Age = 20 **Then** HIV = Positive with plausibility = 0.33333

The final stage would be to test the newly created rules on a test set. The accuracy achieved is then sent back to the GA to optimise the partition sizes.

4 Genetic Algorithm

A genetic algorithm (GA) is a stochastic search procedure for combinatorial optimisation problems based on the mechanism of natural selection [10]. GAs are popular and widely used due to their ease of implementation, intuitiveness and the their ability to solve highly nonlinear optimisation problems. Genetic algorithms are a particular class of evolutionary algorithms that use techniques inspired by evolutionary biology such as inheritance, mutation, selection, and crossover. The fitness/evaluation function is the only part of the GA that has any knowledge about the problem. The fitness function tries

to maximise the AUC. The GA represents the design variables of each individual design with binary strings of 0's and 1's, these are referred to as chromosomes. The variables are limited to the predefined upper and lower bound values. These limits are coded into the GA encoding feature. They perform the task of optimisation, the GA employs three main operators to propagate its population from one generation to the next.

The first operator is *selection*. During each successive generation, a proportion of the population is selected to breed a new generation. This is done on the basis of "survival of the fittest". The fitness of each solution is evaluated using a fitness function, and it is this function that maximises the AUC. After each generation, the AUC is evaluated and sent to the rough set for evaluation, this process continues until the termination criteria is reached. Several selection functions are available and in this paper tournament selection is used. Tournament selection is whereby a "tournament" is run on a few individuals chosen at random and the strongest individual (winner) is chosen for the process of crossover.

The second operator is *crossover* which mimics reproduction or mating in biological populations. The crossover technique used is uniform crossover; in this technique, two parents are combined to produce two new offspring. The way in which it works is individual bits in a string of two parents are compared, and then swapped with a probability of 0.5.

The third operator used is *mutation*, this is whereby an arbitrary bit in a generic sequence or string will be mutated or changed. This is analogous to biological mutation. The reason for implementing this operator is to promote diversity from one generation of chromosomes to another, this prevents the GA from getting stuck at a local optimum but rather a global optimum. Boundary mutation is chosen, this is whereby a variable is randomly selected and is set to either the upper or lower bound depending on a randomly generated uniform number. It must be noted that the chosen operators are case specific, dependent on the nature of the optimisation, different operators would have to be explored.

The pseudo-code algorithm for genetic algorithms is given below;

1. Initialise a population of chromosomes
2. Evaluate each chromosome (individual) in the population
   a) Create new chromosomes by mating chromosomes in the current population (using crossover and mutation)
   b) Delete members of the existing population to make way for the new members
   c) Evaluate the new members and insert them into the population
3. Repeat stage 2 until some termination condition is reached, in this case until 100 generations are reached.
4. Return the best chromosome as the solution

An initial population of 20 individuals is chosen. As mentioned in order to prevent premature convergence to a local minima, the mutation diversification

mechanism is implemented. Other diversification mechanisms such as elitism can also be implemented in an attempt to improve the accuracy.

5 Results Obtained

First the accuracy of the rough set is computed using equal width bin (EWB) partitioning and the resulting accuracy is 69.8%. Next the GA discretisation method is run with GA using a tournament selection function, a boundary mutation and uniform crossover. An initial population of 20 members is selected and the termination function is 100 generations. The GA version produced an accuracy of 87.5%.

Tabulated below are the number of rules (No. Rules), AUC and ROC curves obtained using equal width bin partitioning (EWB) and as well when using a genetic algorithm (GA) to discretise the partitions.

	No. Rules	**AUC**
EWB	488	0.698
GA	49	0.875

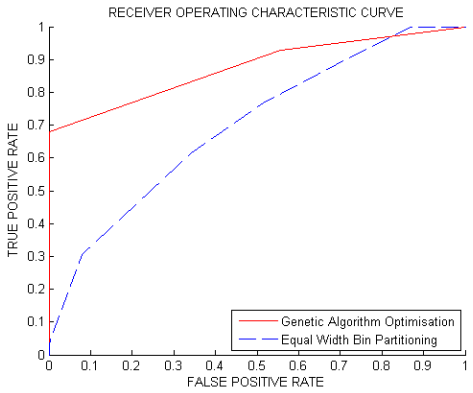

Fig. 1: Tabulated results and ROC curves obtained for EWB and GA optimised partitioning

The ROC curve is chosen to illustrate the results as they allow the performance of the classifiers to be evaluated on how well they predict. The area under curve (AUC) of the respective ROC curves is used as the performance criterion.

Once RST is applied to the HIV data, for the first case of equal width bin partitioning, the 283 cases of the lower approximation are rules that always hold, or are definite cases. The 205 cases of the upper approximation can only be stated with a certain plausibility. For the second case of GA optimised partitions, there are 26 unique discernible cases (in the lower approximation) and 28 indiscernible cases (in the upper approximation). It can be seen that a lower amount of rules resulted in a higher accuracy.

6 Conclusion

A genetic algorithm is successfully applied to RST on the HIV data set. As a result of implementing RST on the data set, the rules extracted are explicit and easily interpreted. RST will however compromise accuracy over rule interpretability, and this is brought about in the discretisation process where the granularity of the variables are decreased. An accuracy of 69.8% is produced by the rough set when applied to the HIV data set for equal width partitioning, an improved accuracy of 87.5% is achieved for the genetic algorithm optimised partitions. Recommendations for future work include the application of other optimisation techniques such as particle swarm optimisation (PSO). PSO is advantageous over GAs as it is easy to implement and there are fewer parameters to adjust.

References

1. A. Lasry, G. S. Zaric, and M. W. Carter. "Multi-level resource allocation for HIV prevention: A model for developing countries." *European Journal of Operational Research*, vol. 180, p. 786799, 2007.
2. "UNAIDS." www.unaids.org/en/HIV_data/2006GlobalReport/default.asp/. Last accessed: 20/3/2007.
3. K. E. Poundstone, S. A. Strathdee, and D. D. Celentano. "The Social Epidemiology of Human Immunodeficiency Virus/Acquired Immunodeficiency Syndrome." *Epidemiol Reviews*, vol. 26, pp. 22–35, 2004.
4. T. Rowland, Ohno-Machado, and A. Ohrn. "Comparison of multiple prediction models for ambulation following spinal cord injury." *In Chute*, vol. 31, pp. 528–532, 1998.
5. Z. Pawlak. *Rough Sets, Theoretical Aspects of Reasoning about Data*, chap. 3, p. 33. Kluwer Academic Publishers, 1991.
6. M. Inuiguchi and T. Miyajima. "Rough set based rule induction from two decision tables." *European Journal of Operational Research*, vol. In Press, Corrected Proof, 2006.
7. R. Department of Health. "National HIV and Syphilis Sero-Prevalence Survey of Women Attending Public Antenatal Clinics in South Africa." http://www.info.gov.za/otherdocs/2002/hivsurvey01.pdf, 2001.
8. A. Ohrn and T. Rowland. "Rough Sets: A Knowledge Discovery Technique for Multifactorial Medical Outcomes." *American Journal of Physical Medicine and Rehabilitation*, vol. 79, pp. 100–108, 2000.
9. F. Witlox and H. Tindemans. "The application of rough sets analysis in activity-based modelling. Opportunities and constraints." *Expert Systems with Applications*, vol. 27, p. 585592, 2004.
10. S. Malve and R. Uzsoy. "A genetic algorithm for minimizing maximum lateness on parallel identical batch processing machines with dynamic job arrivals and incompatible job families." *Computers and Operations Research*, vol. 34, p. 30163028, 2007.