HIV status estimation using optimization, rough sets and demographic data

Rough set theory (RST) is concerned with the formal approximation of crisp sets and is a mathematical tool which deals with vagueness and uncertainty[1,2]. This correspondence presents an approach to optimize rough set partition sizes using various optimization techniques for HIV status estimation. The forecasting accuracy is measured using the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. The four optimization techniques used are genetic algorithm (GA), particle swarm optimization (PSO), hill climbing (HC) and simulated annealing (SA). This proposed method is tested on the human immunodeficiency virus (HIV) data. The results obtained from this granulization method are compared with two previous static granulization methods, namely equal-width-bin (EWB) and equal-frequency-bin (EFB) partitioning.

RST is a mathematical tool which deals with vagueness as well as uncertainty, and it allows for the approximation of sets that are difficult to describe with the available information. It is of fundamental importance to artificial intelligence (AI) and cognitive science, and is highly applicable to the tasks of machine learning and decision analysis. The advantages of rough sets as with many other AI techniques are that they do not require rigid a priori assumptions on the mathematical nature of such complex relationships as do commonly used multivariate statistical techniques[2]. RST is based on the assumption that the information of interest is associated with some information of its universe of discourse[3]. The main concept of RST is an indiscernibility relation (indiscernibility meaning indistinguishable from one another). RST handles inconsistent information.

For knowledge acquisition from data with numerical attributes, a special technique is applied called discretization[4]. Several methods are currently used to perform the task of discretization and these include EWB and EFB partitioning[5]. In this study, the use of combinatorial optimization techniques to discretize the rough set partitions is explored. The four optimization techniques used are GA, PSO, HC and SA[6]. The results produced from these four methods are compared to those of the more commonly used EWB and EFB partitioning methods. To test this method, HIV data are used.

The process of modelling of rough set can be broken down into five stages:

(i) Select the data. The two datasets to be used to test the method are obtained from the South African antenatal survey[7] of 2001.

(ii) Pre-process the data to ensure it is ready for analysis. This stage involves discretizing the data and removing unnecessary data. Although the optimal selection of set sizes for the discretization of attributes is not known at first, an optimization technique is run on the set to ensure that the highest degree of accuracy is obtained when forecasting outcomes.

(iii) If reducts are considered, use the cleaned data to generate reducts[1]. A reduct is the most concise way in which we can discern object classes.

(iv) Extract rules.

(v) Test the newly created rules on a test set.

The methods which allow continuous data to be processed involve discretization, and here EWB and EFB partitioning are investigated. These methods are compared against the proposed method of using various optimization techniques to discretize the continuous data. The optimization techniques are run to create a set of four partitions for the given input data. Using these partitions, the rough set model is generated and the classification accuracy is determined using the AUC of the model produced against the unseen testing data. This result (AUC) is sent back to the optimizer and the partition sizes are changed accordingly to ensure that the rough set produces a better model, i.e. a model with higher classification accuracy. Figure 1 presents a schematic diagram on the process of creating a rough set model.

HIV is well known as being the cause for development of acquired immunodeficiency syndrome (AIDS). In the last 20 years, over 60 million people have been infected with HIV, and among them 95% are in developing countries[8]. The proposed method is tested on demographic data obtained from the South African antenatal sero-prevalence survey of 2001. The amount of data used is 12945. The dataset was balanced and then split into training and testing data using the ratio of 70 : 30% respectively. The performance of the rough set model was validated using the testing data. 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 results of EWB and EFB partitioning are compared to those of the various optimization approaches. The GA was run with roulette wheel selection, a boundary mutation and uniform crossover, an initial population of 20 individuals and the termination function of 100 generations. PSO was implemented with the maximum number of generations of 30, and the initial number of particles also set to 30. Steepest ascent HC was implemented with 20 initial starting points. SA was run with a random generator with the bounds of the maximum and minimum input values, an initial temperature of 1, a stopping temperature of $1e^{-8}$, a maximum number of consecutive

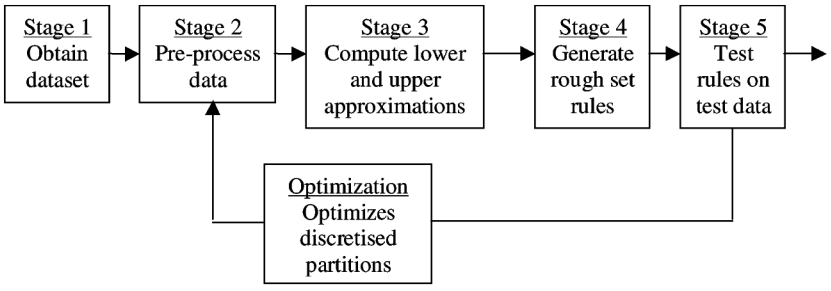

Figure 1. Schematic diagram on the process of creating a rough set model.

Table 1. Results obtained for the four optimization and two static discretization methods

	Number of rules	Computation time (s)	AUC
Genetic algorithm	172	3726	0.6748
Particle swarm	146	1690	0.6718
Hill climbing	209	12624	0.6902
Simulated annealing	197	1159	0.6802
Equal width bin	231	2	0.5952
Equal frequency bin	307	2	0.5986

rejections of 200 and a maximum number of successes within one temperature set to t_0.

Table 1 shows that, albeit marginal, HC produces the highest classification accuracy when using the AUC measure. SA optimization has the lowest computational time with respect to the optimized methods. From Table 1, it is evident that the computational time required for the static EWB and EFB partitioning is much less than that for the optimized approaches. Having said that, the higher forecasting accuracy obtained using optimized methods results in more concrete evidence from which policies can be generated. Fewer rules are also generated from the optimized methods approach, and it is from these extracted rules that the causal interpretations are then formulated by a linguistic approximation.

The results indicate that the proposed method of using optimization techniques to granulize/discretize rough set partitions is feasible and it also produces higher forecasting/classification accuracies than the EWB and EFB partitioning. Among the four optimization techniques, it can be stated that no particular technique is superior to another. There are marginal differences in the accuracies produced (using AUC), but in all cases easy-to-interpret, linguistic rules are generated.

The results of both datasets indicate that the optimized methods produce higher forecasting/classification accuracies than the static EWB and EFB partitioning methods. When comparing the optimized approaches against each other, as expected there is no significant difference between the methods used. The rough sets produce a balance between transparency of the rough set model and accuracy of HIV estimation, but it does come at a cost of high computational effort.

1. Pawlak, Z., *Rough Sets, Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, 1991, p. 33.
2. Garson, G. D., *Soc. Sci. Comput. Rev.*, 1991, **9**, 399–433.
3. Komorowski, J., Pawlak, Z., Polkowski, L. and Skowron, A., *The Handbook of Data Mining and Knowledge Discovery*, Oxford University Press, 1999.
4. Grzymala-Busse, J. W., In Proceedings of the Rough Sets and Emerging Intelligent Systems Paradigms, June, 2007, pp. 12–21.
5. Jaafar, A. F. B., Jais, J., Hamid, M. H. B. H. A., Rahman, Z. B. A. and Benaouda, D., In Proceedings of the 4th International Conference on Multimedia and Information and Communication Technologies in Education, Seville, Spain, November 2006.
6. Marwala, T., *Computational Intelligence for Modelling Complex Systems*, Research India Publications, Delhi, 2007.
7. Crossingham, B. and Marwala, T., In *Studies in Computational Intelligence*, Springer-Verlag, 2007, vol. 78, pp. 245–250.
8. Lasry, A., Zaric, G. S. and Carter, M. W., *Eur. J. Oper. Res.*, 2007, **180**, 786–799.

Received 31 March 2008; revised accepted 24 September 2008

TSHILIDZI MARWALA*
BODIE CROSSINGHAM

School of Electrical and Information Engineering,
University of the Witwatersrand,
Private Bag X3, WITS,
2050, South Africa
**For correspondence.*
e-mail: tshilidzi.marwala@wits.ac.za

Occurrence of zincian ilmenite from Srikurmam placer sand deposit, Andhra Pradesh, India

Placer deposits of Andhra Pradesh (AP) particularly ilmenite occur at Bhavanapadu, Kalingapatnam, Srikurmam and Donkuru-Barua in Sikakulam District, Bhimunipatnam in Visakhapatnam–Vizianagaram districts, Kakinada in East Godavari District and Nizamapatnam in Guntur and Prakasham districts. Srikurmam mineral sand deposit is one of the largest placer deposits established from this region (A. Y. Rao *et al.*, unpublished). The deposit is confined between two prominent lineament-controlled rivers, Nagavali and Vamsadhara, and spans over a coastal length of 22 km. Geomorphic units are dominated by the presence of structurally controlled estuarine rivers, viz. Nagavali in the south and Vamsadhara in the north. A tidal creek and a geomorphic-low, lppligedda, occur to the north-central part of the area (Figure 1). The fore dune and rear dune are well developed, whereas inter dune is masked at places. The dunes generally rise to a height of 18 m in the rear, as a consequence of resting on higher basement of palaeo-dunal-beach complex (Figure 2). It is a general assumption that the source of heavy minerals for the deposit is the granulite facies rocks of Khondalite group. Charnockite has a restricted occurrence in the hinterland, dominated by Khondalite. Upper Gondwana formations occur to the south-west of the deposit. The area also receives its detrital material from the older reddened dunes exposed along the western margins and the offshore bars. Preliminary investigations have revealed a total mineral content of 30 million tonnes at a working grade of 34.36%, and is by far the richest deposit in this part of the coast. Marginally higher garnet content (37.10%) has been estimated along with an ilmenite content of 31.94% (A. Y. Rao *et al.*, unpublished). Ramana (unpub-