NEW COMPACT ROUGH CLASSIFICATION MODEL

Walid Saeed*, Md Nasir Sulaiman, Mohd Hasan Selamat, Mohamed Othman and Azuraliza Abu Bakar

Recived: Dec 12, 2003; Revised: Jul 2, 2004; Accepted: Jul 7, 2004

Abstract

This article deals with rough classification mining. It presents a strategy on knowledge discovery in the Information Systems (IS) based on rough set approach. It also presents the Effective Integral Programing (EIP) model in data mining rough classification modeling. The model is based on generating a 0-1 integer programing model from rough discernibility relations of a decision system (DS) to get minimum selection of significant attributes, which is called reduct in rough set theory. New algorithms in the searching process proposed to solve the EIP model are called Extracting Effective Rules (EER) algorithms. The experiments on sets of data show that the EIP model has good accuracy and the proposed EER algorithms have reduced the number of rules generated from the EIP model.

Keywords: Data mining, rough set, decision system, reduct

Introduction

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information, which can be used to increase revenue, cut costs, or both. It has gained considerable attention among practitioners and researchers as evidenced by the number of publications, conferences, and application reports (Saeed *et al.*, 2003b; Saeed *et al.*, 2003c). The growing volume of data that is available in a digital form has accelerated this interest. Data mining relates to other areas, including machine learning, cluster analysis, regression analysis, and neural networks (Kusiak, 2001). Data mining researchers often use classifiers to identify important classes of objects within a data repository. Classification is particularly useful when a database contains examples that can be used as the basis for future decision-making. Although the classification is an important and useful process in knowledge representation systems, the processing time increases rapidly as the size of the knowledge base increases (Kim, 1993). The objective of this study is to present the EIP model in data mining rough classification, and EER algorithms to solve the EIP model. The paper is structured as follows. Related work is briefly explained in section 2. The EIP model is described in section 3. The Extracting Effective Rules algorithms and selected data sets are described in sections 4 and 5 respectively. Experimental results and the conclusion are presented in sections 6 and 7.

Faculty of Computer Science and Information Technology, University Putra Malaysia 43400, UPM Serdang, Selangor, Malaysia, E-mail: walidsh2001@yahoo.com

[*] *Corresponding author*

Related Work

In this section four selected classification algorithms that are related to the proposed approach are briefly described.

(SIP/DRIP) Algorithm

The algorithm Standard Integer Programing (SIP) / Decision Related Integer Programing (DRIP) transforms the discernibility relations from the equivalence class into an IP model (Bakar *et al.*, 2001a). SIP model is used to find minimal reducts of each class in the equivalence class and the DRIP model is used to find the minimal reduct of the whole DS (Figures 1, 2), which is called reduct in rough set theory (Bakar *et al.*, 2001b; Bakar, 2001c).

```
Input: An Equivalence Class Ei,
Output: An IP Model
j= i +1; //i, j: class number
while (j < total class)
{ for (k = 0; k <num attribute; k++)
{ if ak(Ei) _= ak(Ej)
mik = 1
else
mik = 0
}
}
```

Figure 1. SIP algorithm.

```
Input: An Equivalence Class Ei,
Output: An IP Model
j= i +1; //i, j: class number
while (j < total class)
{ for (k = 0; k <num attribute; k++)
{ if ak(Ei) _= ak(Ej) and δk(Ei) _= δk(Ej)
mik = 1
else
mik = 0
}
}
```

Figure 2. DRIP algorithm.

Genetic Algorithm

Genetic algorithm is an iterative procedure (Figure 3) that consists of a constant-size population of individuals, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space (Bari, 2001). This space, referred to as the search space, comprises all possible solutions to the problem (Michael *et al.*, 1997). Outline of the basic Genetic algorithm proceeds as follows:

1. [Start] Generate random population of n chromosomes (suitable solutions for the problem)
2. [Fitness] Evaluate the fitness f(x) of each chromosome x in the population
3. [New population] Create a new population by repeating the following steps until the new population is complete
   - [Selection] Select two parent chromosomes from a population according to their fitness (the better the fitness, the bigger the chance to be selected)
   - [Crossover] With a crossover probability cross over the parents to form new offspring (children). If no crossover was performed, any offspring is an exact copy of parents.
   - [Mutation] With a mutation probability mutate new offspring at each locus (position in chromosome).
   - [Accepting] Place new offspring in a new population
4. [Replace] Use new generated population for a further run of the algorithm
5. [Test] If the end condition is satisfied, stop, and return the best solution in the current population
6. [Loop] Go to step 2

Johnson Reducer

The Johnson Reducer algorithm invokes a simple greedy algorithm to compute a single reduct only. Let fA denote a suitably constructed discernibility function. The reduct B is then found as follows:

1. Initialize B to the empty set
2. While the function fA has any sums left, do the following:

```
// start with an initial time
t := 0;
// initialize a usually random population  of individuals
initpopulation P (t);
// evaluate fitness of all initial individuals of population
evaluate P (t);
// test for termination criterion (time, fitness, etc.)
while not done do
    // increase the time counter
    t := t + 1;
    // select a sub-population for offspring production
    P' := selectparents P (t);
    // recombine the "genes" of selected parents
    recombine P' (t);
    // perturb the mated population stochastically
    mutate P' (t);
    // evaluate its new fitness
    evaluate P' (t);
    // select the survivors from actual fitness
    P := survive P, P' (t);
od
end GA
```

Figure 3. Genetic algorithm.

- Let a denote the attribute that maximizes ∑w(s), where s occurs in fA and a occurs in s
- Add a to B
- Delete all sums from fA that contain a.

where w(s) denotes a weight for sum s in fA that automatically is computed from the data.

Holte1R Reducer

The 1R procedure for machine learning is a very simple one that proves surprisingly effective on the standard data sets commonly used for evaluation. 1Rs are rules that classify an object on a basis of a single attribute that takes a set of training examples as input, each with several attributes and a class and a 1-rule output. The aim is to infer a rule that predicts the class given the values of the attributes. The 1R algorithm chooses the most informative single attribute and bases the rule on this attribute alone. The algorithm suggests that it may be possible to use the performance of 1-rules to predict the performance of the more complex hypotheses produced by standard learning systems (Craig *et al*., 1995). The following algorithm (Figure 4) shows the basics of Holte1R procedure.

```
For each attribute a, form a rule as follows:
    For each value v from the domain of a,
        Let c be the most frequent class in the set of
            instances where a has value v.
        Add the following clause to the rule for a:
            if a has value v then the class is c
    Calculate the classification accuracy of this rule.
Use the rule with the highest accuracy.
```

Figure 4. Holte1R reducer algorithm.

EIP Model

The EIP model is based on generating 0-1 values (Figure 5) from rough discernibility relations of a DS in order to get the minimum selection of important attributes, which is called the reduct in rough set theory. The idea of the model is generating value one when the attributes values of two classes are different or when the decision is that the values are the same, otherwise value zero is generated. There are two algorithms used (Saeed *et al.,* 2003a) in order to solve the EIP model to obtain the full reduct of the DS and the rules of the classes.

```
Input:    An Equivalence Classes Ei
Output:   EIP Model

j = i +1;                          // i, j : class number
while (j < m )  {          // m : number of classes
     for (k = 0; k < n; k++) {
                               // n : number of attributes
          if  $a_k (E_i) \neq a_k (E_j)$ or $d_k$ (Ei)  =  $d_k$  (Ej)
                                    $b_{ik} = 1$
                       else
                                    $b_{ik} = 0$
}
  }
```

Figure 5. Effective integer programing algorithm.

Extracting Effective Rules (EER) Algorithms

Two algorithms are proposed to solve the EIP model. The first is called Extracting All Rules (EAR), which examines the EIP model to find all rules in the DS which exactly represent all the decision system. The second is called Extracting Full Reduct (EFR), which examines the DS to find the full reduct of the DS.

Extracting All Rules Algorithm

This algorithm examines the EIP model class by class to find all effective rules in the DS for every class. All these rules exactly represent the DS. The EAR algorithm is shown in Figure 6.

```
Input:    EIP model
Output: Effective Rules for every class

For ( cl = 0; cl < Class_No; cl++ )
     For ( xv = 0; xv < Attribute_No ^ 2-2 ; xv++ )
          {     z_lower=0;
                 For ( j =0; j < EIP_Class; j++)  // EIP_Class:
                           // number of  EIP for every classes
                   {   z_lower = Check_Value()
                      If ( z_lower = 0 )
                        Break
                   }
                 z_upper = Calculate_Value()
                 if ( z_upper < Attribute_No)
                        Add_New_Rule()
          }
```

Figure 6. Extracting all rules algorithm.

Extracting Full Reduct Algorithm

This algorithm examines all the EIP model as one group to find the full reduct of the DS, which means that those attributes can represent the DS. The EFR algorithm is shown in Figure 7.

Table 1 shows five equivalence classes of 100 objects. EIP model is obtained from Table 1 for every class. For example the EIP model for class 2 is shown in Table 2. When EAR is applied on the EIP model of class 2 the obtained rules are:

a2 → d2
a2b2 → d2
a2c3 → d2

When EFR is applied on the EIP for all classes the full reduct is (a, b) which means that just two attributes can represent all the DS.

Selected Data Sets

Four data sets are selected and applied in our study. The data sets are Australian Credit Card Approval *(AUS)*, Cleveland Heart Disease *(CLEV)*, Lymphography *(LYM)* and Breast Cancer *(BCO)* data sets. These data sets were chosen to evaluate the selected algorithms capabilities under controlled conditions for specific data characteristics. The data sets were drawn from the UCI-Irvine repository of machine learning databases (Murphy, 2002). Some characteristics of these data sets are shown in Table 3.

Experimental Results

In this section the results of several practical experiments to examine the performance of

Table 1. Equivalence classes of 100 objects.

Class	Attributes			Decision
	A	b	c	
E0	1	2	3	1 (50 x)
E1	1	2	1	2 (5 x)
E2	2	2	3	2 (30 x)
E3	2	3	3	2 (10 x)
E4	3	5	1	3 (5 x)

Table 2. EIP model for class 2.

1	0	0
1	0	1
0	1	0
1	1	1

```
Input:    EIP model
Output:   Full reduct for all system

For ( xv = 0;  xv < Attribute_No ^ 2-2 ; xv++ )
{    z_lower=0;
     For ( j =0 ;j < EIP_no ; j++)
            // EIP_no: classes number in EIP
       {  z_lower= Check_Value()
          If ( z_lower = 0 )
             Break
       }
        z_upper = Calculate_Value()
        if ( z_upper < Attribute_no)
               Add_New_Reduct()
 }
```

Figure 7. Extracting full reduct algorithm.

Table 3. Characteristics of the selected datasets.

Domain	Set size	# & Type of features			# of classes
AUS	690	6 C,	9 D	(15)	2
CLEV	303	6 C,	8 D	(14)	2
LYMP	148	3 C,	15 D	(18)	4
BCO	699	9 C		(9)	2

C = Continuous, D = Discrete

Table 4. The classification accuracy.

Data	Methods				
	EIP	SIP/DRIP	GA	Johnson	Holte1R
AUS	94.29	85.37	81.60	76.40	67.95
CLV	83.25	79.60	81.13	75.92	69.30
LYM	88.10	82.16	81.60	75.84	65.45
BCO	88.33	89.95	94.06	93.16	77.86

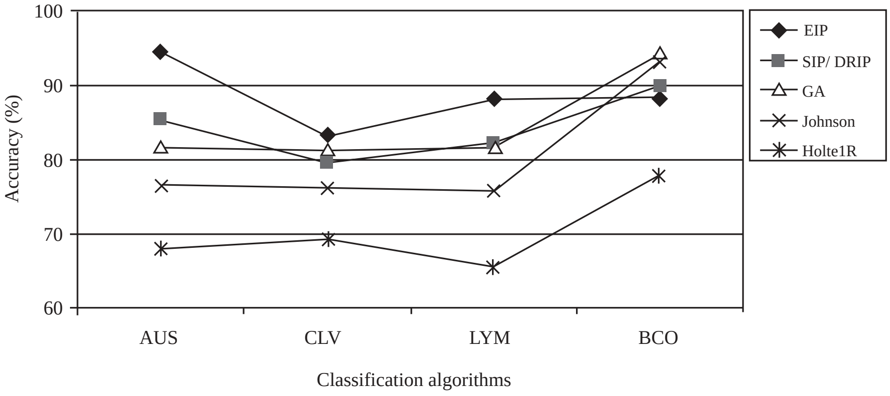

Figure 8. Classification accuracies comparison of classification algorithms.

different types of algorithms on real world problems are presented. All experiments were carried out on four data sets obtained from the UCI repository and compared and applied on four methods:- SIP/DRIP, Genetic algorithm, Johnson algorithm and Holte1R algorithm. The results in Table 4 and Figure 8 show that the EIP model provides good classification as compared with other methods. Especially, the EIP model was the best method on the three testing data sets *AUS*, *CLEV* and *LYMP*. We note that some algorithms are good with some data sets, but are not effective with others; which means that the effectiveness of an algorithm depends on the nature and type of the data sets.

Conclusion

This paper discussed the proposed Effective Integral Programing model in finding interesting pieces of knowledge from the decision system. The Effective Integral Programing model is implemented within the rough set framework in generating rules. The experimental results indicate that the rules generated from the proposed reducts calculation method have given a good classification model with good classification accuracy. This shows that the proposed Effective Integral Programing model rough method has generated a good selection of knowledge from the decision system and the model is able to perform well with different data sets.

References

Bakar, A.A., Sulaiman, M.N., Othman, M., and Selamat, M.H. (2001a). IP algorithms in compact rough classification modeling. Intelligent Data Analysis, IOS Press, Amsterdam, 5(4):419-429.

Bakar, A.A., Sulaiman, M.N., Othman, M., and Selamat, M.H. (2001b). Improved rough classification model: A comparison with neural classifier. Journal Institute of Mathematics & Computer Science (Comp.

Sc. Series), 12(1):37-46.

Bakar, A.A. (2001c). Propositional satisfiability method in rough classification modeling for data mining, [Ph.D. thesis]. Computer Science, Computer Science and Information Technology, University Putra Malaysia, Malaysia, 221 p.

Bari, P.D. (2001). Simulated annealing vs. genetic algorithms for linear spline approximation of 2D scattered data. Proceedings of the XII ADM International Conference; Sept 5-7, 2001; Grand Hotel, Rimini, Italy, 9 p.

Kim, J. (1993). Classification and retrieval of knowledge on a parallel marker-passing architecture. IEEE Transactions on Knowledge and Data Engineering, 5(5): 753-761.

Kusiak. (2001). Rough set theory: A data mining tool for semiconductor manufacturing. IEEE Transactions on Electronics Packaging Manufacturing, 24(1):44-50.

Murphy, P.M. (2002). UCI Repositories of Machine Learning and Domain Theories [URL]. Available from: http://www.isc.uci.edu/~mlearn/MLRepository.html. Accessed Nov 22, 2002.

Nevill-Manning, C.G., Holmes, G., and Ian, H. (1995). The development of halite's 1R classifier, Los Alamitos, CA. Proc. Artificial Neural Networks and Expert Systems. IEEE Computer Society Press, Dunedin, New Zealand, 4,273(6):5,250.

Raymer, M., Punch, W., Goodman, E., Sanschagrin, P., and Kuhn, L. (1997). Simultaneous feature extraction and selection using a masking genetic algorithm. Proceedings of the 7th International Conference on Genetic Algorithms; July 19-23, 1997; Morgan Kaufmann Publishers; San Francisco, East Lansing, MI, USA, 7 p.

Saeed, W., Sulaiman, M.N., Selamat, H., Othman, M., and Bakar, A.A. (2003a). EER algorithm to solve IP model in compact rough classification modeling. Proceedings of the Malaysian-Japan Seminar on Artificial Intelligence Application in Industry; Jun 24-25, 2003; University Technology Malaysia, International Park Plaza Hotel, Kuala Lumpur, Malaysia, 7 p.

Saeed, W., Shiba, O., and Sulaiman, M. (2003b). Comparative study of data mining classification algorithms. Proceedings of the Advanced Technology Congress; May 20-21, 2003; Putrajaya Marriot Hotel, IOI Resort, Putrajaya, Institutes of Advance Technology, University Putra Malaysia, Malaysia, 8 p.

Saeed, W., Shiba, O., Sulaiman, M.N., Mamat, A., Selamat, H., Othman, M., Ahmed, D.F., and Bakar, A.A, (2003c). Classification algorithms: experiments and comparative study. Proceedings of the Integrating Technology in the Mathematical Sciences; April 14-15, 2003; University Science Malaysia, Vistana Hotel, Benang, Malaysia, 7 p.