Rough Sets for Mining Educational Data

Julia Ann Johnson
Computer Science
University of the Laurentian
Sudbury, Ontario, Canada
jjohnson@ cs.laurentian.ca

Genevieve Marie Johnson
School of Education
Curtin University
Perth, Western Australia
g.johnson@curtin.edu.au

Robert F. Cavanagh
School of Education
Curtin University
Perth, Western Australia
r.cavanagh@curtin.edu.au

Abstract **-** While educational data are typically analyzed with statistical software, data mining techniques are increasingly appropriate in revealing complex relationships among multiple variables in large amounts of data. We experimented with the rough set method in conjunction with statistical analysis to identify patterns in, and thereby extract meaning from, complex educational data. Results establish the benefits of combining rough set decision making with stochastic analysis in mining exceedingly complex and difficult to interpret educational data sets.

Keywords - uncertain reasoning, rough sets, educational data, data mining

I. INTRODUCTION

School-based learning is a complex phenomenon because a wide range of student, teacher, parent and classroom variables combine to impact on the quality and quantity of academic achievement [1]. Given that all children are required to attend school and that schools are supported entirely by public funds, a considerable volume of educational data are available. While a statistical approach based on SPSS is prevalent for analyzing social sciences data [2], such methods do not consistently reveal relationships that we know intuitively exist in the data. To confirm our intuitions we look to data mining techniques for revealing relationships between variables that have remained hidden using traditional statistical techniques. Furthermore, traditional techniques have sometimes revealed unusual findings that are inconsistent with expectations. Does data mining reveal the same unexpected outcomes on the same data set? When traditional statistical techniques have uncovered no information, does data mining also fail at finding any?

II. DATA MINING TECHNIQUE

Data mining and data warehousing were researched by the authors [3, 4, 5, 6] and rudimentary applications on a variety of domains in computer science, engineering and the social sciences were implemented. Those applications used mining and warehousing software based on rough sets.

A. Rough Set Paradigm

The notion of *Rough Set* underlies many algorithms for reasoning with uncertain or vague data presented in tables. The uncertainty comes from our inability to distinguish objects from each other arising because objects cannot be completely specified by their known properties. Rough set theory provides a framework for dealing with such type of uncertainty. An indefinable set is approximately represented by two definable sets, called lower and upper approximations.

B. Decision Making Software

Reverse prediction method [4] enables prediction of the values for condition attributes that imply given values for the decision attributes. Ordinary prediction is expressed as follows where C_i are condition attributes and D are decision attributes:

$$C1[\text{given}], C2[\text{given}], \ldots, Cn\ [\text{given}] \rightarrow D[\text{predict}]$$

It is possible to reverse the roles played by the condition and decision attributes while still employing ordinary prediction:

$$D[\text{given}] \rightarrow C1[\text{predict}], C2[\text{predict}], \ldots, Cn[\text{predict}]$$

It is also possible to change their role as given or predicted:

$$C1[\text{predict}], C2[\text{predict}], \ldots, Cn[\text{predict}] \rightarrow D[\text{given}]$$

Rough Set Reverse Prediction Algorithm (RSRPA) was implemented and embedded in an interface called Rough Set Graphical User Interface (RSGUI). The capability to reverse the roles of attributes figures significantly in the provision of useful tools for defining concepts implicitly described in tabular data.

III. DATA DESCRIPTION

A survey was developed that asked secondary school students to rate their level of agreement (scored '1' for disagree, 2' for agree and '3' for strongly agree) with 85 statements. The statements assessed a range of educationally relevant characteristics including students' evaluation of themselves and their abilities, their perception of classroom requirements, the value they attached to education, learning outcomes, and perception of teacher and parent support for their learning.

Data were collected from August 2010 to December 2010. 4500 surveys were distributed to students in 23 schools in Perth, Western Australia and 1760 (39%) were returned and processed. Approximately 46% of the surveys were completed by boys and 56% were completed by girls. Approximately 22% of surveys were completed by students in Grade 8, 20% in Grade 9, 30.4% in Grade 10 and 27.8% in Grade 11.

A sample of statements labeled SE1, SE2, et cetera is illustrated in Fig. 1. The statement labels reappear as column

names in an excel spread sheet that houses the data. Each row corresponds with a rating of 1 to 3 for each statement by each respondent.

Statement labels encode to some degree the relationships among columns. The self-esteem statements SE1, SE2 and SE5 are expected to be correlated with each other because they are measuring the same construct (i.e., self-esteem). For a given individual, a rating for "I am OK" is expected to have a similar rating as that for "I am pleased with myself". Likewise, statements R2, R3 and R5 are expected to be correlated because they all reflect aspects of student resilience.

Columns were partitioned into three different sections or Parts A, B and C. Statements in Part A reflect psycho-educational attributes conceptualized as within-student varieties (i.e., self-esteem, resilience, self-regulation and self-efficacy). Statements in Part B reflect student perception of class requirements including *explanation* (e.g., In this class, I am expected to connect different ideas together), *interpretation* (In this class, I am expected to show I know the work correctly), *application* (In this class, I am expected to practice using what I've learnt), *perspective* (In this class, I am expected to think about the views of experts when I am learning new things) and *empathy* (In this class, I am expected to try to understand the views of others). Part C and its subgroups (C1 to C8) measure multiple aspects of classroom learning environments including student self-report of *educational values* (I enjoy finding out how things work), *learning outcomes* (I understand the work well), *classroom learning* (Students learn from each other), *classroom support*, (Students support each other), *classroom discussion* (We talk about our progress), *classroom planning* (We are involved in deciding how our progress will be assessed), the *teacher* (The teacher asks our advice) and *parents* (My parents take an interest in my progress).

Multiple regression analyses were conducted on a subset of these data [7]. Multiple regression analysis, a common approach to analysing educational data, is a technique for modelling and analysing several variables. Specifically, regression analysis describes the extent to which the dependent variable (for example, student examination results) changes when an independent variable (for example, student score on a measure of self-esteem) is varied [8]. The classroom learning attitudes and behaviours of students were found to relate directly to educational outcomes, as were teacher expectations and parent attitudes and behaviours. The attitudes and behaviours of students and teachers towards classroom collaboration and caring were not confirmed to relate directly to learning outcomes. That is, student learning (dependent variable) did not change with student perception that their classmates and teacher cared about them. Such a lack of statistical significance is inconsistent with common theoretical assumptions and empirical findings [9, 10, 11]. The question becomes, do data mining techniques applied to these educational data provide information beyond that gleaned via statistical analysis such as multiple regression?

In this class …	
SE1	I am OK
SE2	I am pleased with myself
SE5	I am confident about my ability to perform well
R2	I can overcome small problems
R3	I don't admit defeat easily
R5	Big challenges bring out the best in me
SR1	I make an effort
SR2	I am clear about my strengths and weaknesses
SR4	Improvements in my learning come from me

Figure 1. Sample of survey items

IV. RESULTS

We have experimented with applying a rough sets data mining method to look for patterns in the data described in Section III. The variables were reduced to those 16 survey items that revealed no useful information using regression analysis despite the assumption of relationships between variables [7]. The choice is arbitrary as to which attributes are to be considered as condition and which decision. Nine of the 16 variables measured learning outcomes which were considered to be decision attributes. For the purpose of illustration, only one learning outcome (LO1) was considered in the data mining experiment. The seven remaining variables were taken as condition attributes. Fig. 2 shows the first few rows of a 1496 row table that was analyzed.

Missing values are indicated with a '0' in the table. Rows that contained all zeros were omitted from the data set, the rational being that no information is available from a row with no values. In general, consideration of only a subset of columns of the original data results in a great deal of uncertainty. The remaining columns may result in a row with all zero entries or with all values equal, say to 3 for example. Perhaps such items were carelessly answered by students and thus should be omitted from the data set. Alternatively, there may have been meaningful answers in the fields that have been projected away and the row should be retained.

The tabs along the top of the Fig. 2 show that RSGUI supports a variety of methods available for data analysis (e.g., RS1, RSRPA).

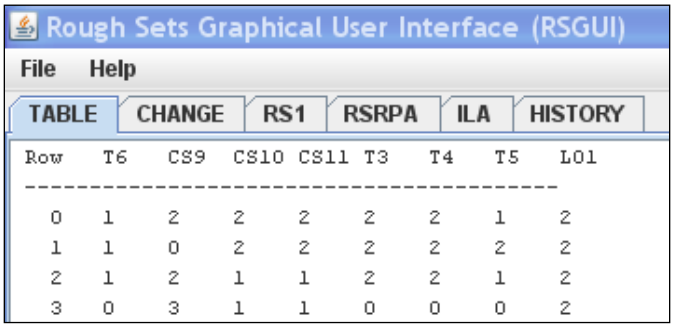

Figure 2. Sample rows of table input to RSGUI

For a given value of the decision attribute, the condition attribute values that lead to that particular decision (concept or outcome) were predicted. Fig. 3 demonstrates execution of RSRPA. The RSRPA tab is highlighted to indicate that it has been made current. RSRPA generated values for the condition variables that would lead to the given concept. Definition of two different concepts is illustrated in the status window. For concept LO1 = 0, there is strongest evidence that the values of the other variables will be as shown in the status window of Fig. 3. The results for all four possible concepts are illustrated in Table 1.

Since a value of zero indicates missing data, predicting LO1 = 0 is subject to interpretation. LO1 = 1 describes the responding secondary students' disagreement, LO1 = 2 describes agreement and LO1 = 3 describes strong agreement with the survey item "I understand the work well." In light of

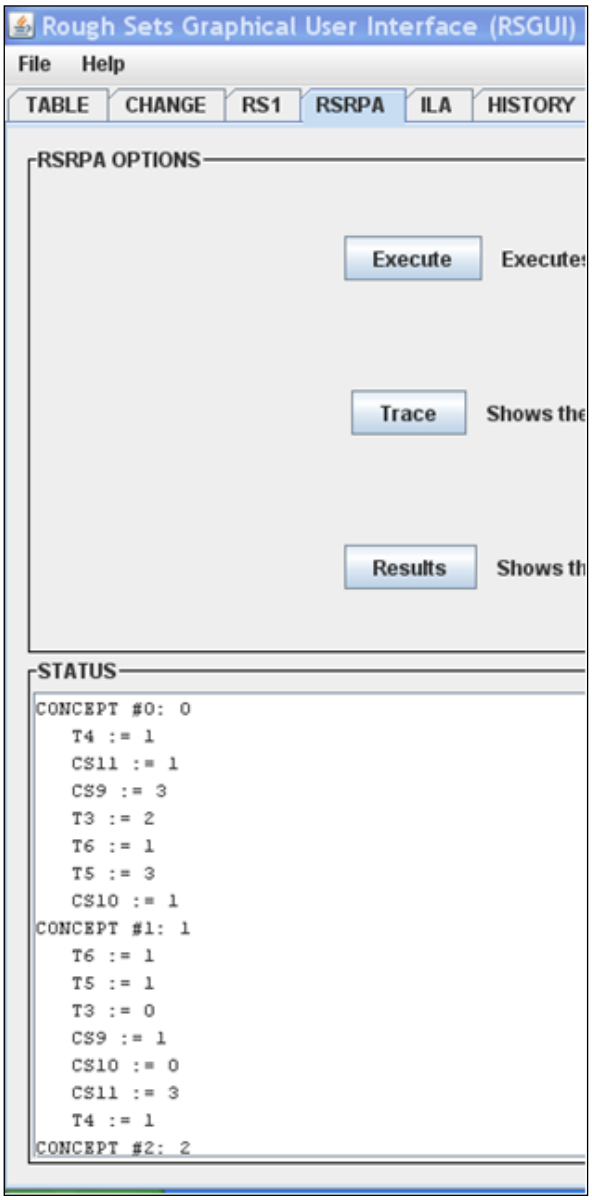

Figure 3. Execution of RSRPA

TABLE I. DEFINITION OF CONCEPTS

	LO1 = 0	LO1 = 1	LO1 = 2	LO1 = 3
CS9	3	1	3	3
CS10	1	0	0	1
CS11	1	3	1	2
T3	2	0	1	3
T4	1	1	0	2
T5	3	1	0	3
T6	1	1	0	3

the meaning of the data summarized in Table 1 and for purposes of illustration, students who claimed that they did not understand their school work also disagreed with the survey items CS9 (Students are tolerant of one another), T4 (The teacher helps students who get into trouble around the school), T5 (The teacher helps students with family problems) and T6 (At times, the teacher seems more like a mum or a dad than a teacher) and strongly agreed with CS11 (Students are not nasty towards each other). That is, unlike multiple regression analysis, application of RSRPA clearly established relationships between student's perception of their own capacity to learn and level of perceived classroom support from peers and the teacher. RSGUI findings are both intuitively and empirically [9, 10, 11] more meaningful than conventional statistical analysis. Establishing the validity of our data mining technique, students who expressed the perception that they understood their school very well (strongly agreed with the survey item) also agreed or strongly agreed with all survey items that indicated teacher support for students and most items that indicated peer support of learning. There was one exception, item CS10 (Students care for each other).

In fact, not all of the given attributes may be required to define a given concept. The RS1 algorithm in contrast with RSRPA, omits variables that provide no new information over that provided by the remaining condition variables. RS1 is an inductive learning algorithm that generates predictive rules using forward (not reverse) prediction. The antecedent conditions of a predictive rule refer to variables that act as predictors and the consequents to outcome variables.

The user clicks on a value (0, 1, 2, 3) to set condition variables one at a time as demonstrated in Fig. 4. The Next>> button results in a pop up as shown for the next condition variable until a value for the last condition variable has been entered when the only available button is predict.

As illustrated in the status window of Fig. 5, a minimal set of variables and the values for defining the concept LO1 = 1 was found to be {T6 = 1, T5 = 1, T3 = 0, CS9 = 1}. These variable-value pairs appear in the status window as antecedent conditions of the strongest rule consistent with the values input for all of the variables as shown at the top of the status

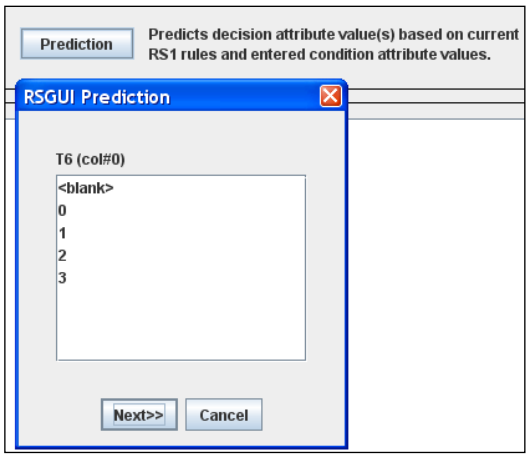

Figure 4. Prediction of decision values

window of Fig 5. That is, students who claimed that they did not understand their school work also disagreed with survey items T6 (At times, the teacher seems more like a mum or a dad than a teacher), T5 (The teacher helps students with family problems) and CS9 (Students are tolerant of one another). The minimal sets are not unique as a different set, may also define the given concept. However, a minimal set contains no redundant attributes.

A rule for predicting LO1 = 2 is also applicable given that particular combination of values for all of the variables. In this case, however, a shorter rule is adequate with antecedent conditions (T6 = 1, CS9 = 1, CS10 = 0). The results are confounded by the missing values for the values entered for prediction. When the missing values were entered as <blank>, no applicable rules were found, indicating that the missing variables are being taken as relevant markers though a human interpretation of their relevance is questionable.

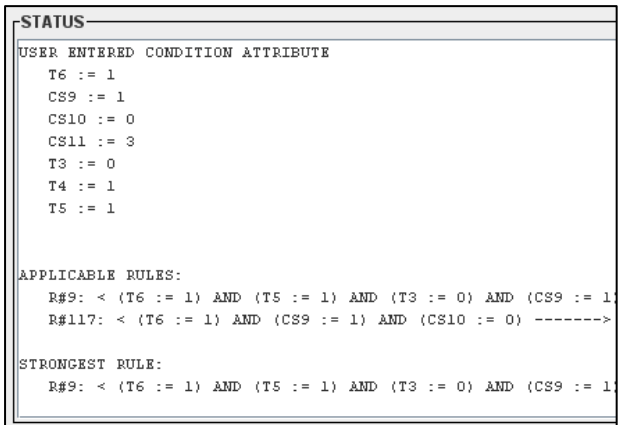

Figure 5. Prediction of decision values

V. CONCLUSION

The same data have yielded both stochastic and deterministic interpretations. Future work is expected in which the scope of the present study is expanded to include the development of criteria for examining educational data for its suitability for mining and warehousing. Traditional database systems with a query capability may also help to manage, integrate and find meaning in the huge stores of data from multiple sources currently held by educational researchers including government Departments of Education and Ministries of Learning.

For future data acquisition, a front end is recommended to ensure that the questionnaires were answered with care. If the data had come from a relational database in the first place, the database management system would have ensured that a certain set of predefined constraints were enforced on the data The complex educational data set examined in this research appears to include a great deal of noise. The Rough Set model is intended for noisy data but the data at hand were further complicated by the presence of unanswered questions. Null values in the data add more possible values for variables without added more meaning thereby decreasing plausibility of the rules generated. Further work is expected in which methods for dealing with incomplete data [12] are introduced to the rule generation process.

The rough set method helped to identify variables that are most important to the knowledge represented in the data. A subset of table attributes was found to, by itself, fully characterize a learning outcome.

REFERENCES

[1] G. M. Johnson, “Internet use and child development: The techno-microsystem,” Australian Journal of Educational and Developmental Psychology, vol. 10, 2010, pp. 32-43.

[2] G. M. Johnson and A. J. Howell, “The impact of Internet learning technology: Experimental methods of determination,” in Selected Styles in Web-Based Educational Research, B. L. Mann, Ed. Hershey, PA: Idea Group Publishing, 2006, pp. 282-301.

[3] J. A. Johnson and G. M. Johnson, “Building knowledge around complex objects using infobright data warehousing technology,” International Journal of Database Theory and Application, vol. 3, 2010, pp. 31-46.

[4] J. A. Johnson and G. M. Johnson, “RSGUI with reverse prediction algorithm,” in Studies in Fuzziness and Soft Computing, vol. 224, R. Bello, R.. Falcan, W., Pedrycz and J. Kacprzyk, Eds. Berlin: Springer, 2008, pp. 287-306.

[5] R. A. Frutuoso-Barroso, G. Baiden and J. Johnson, “Knowledge representation and expert systems for mineral processing using Infobright,” Proceedings of the IEEE International Conference on Granular Computing, 2010, pp. 49-54.

[6] J. A. Johnson and G. M. Johnson, “InfoBright for analyzing social sciences data,” in Communications in Computer and Information Science, vol. 64, D. Ślęzak, T. Kim, Y. Zhang, J. Ma and K. Chung, Eds. Berlin: Springer, 2009, pp. 90-98.

[7] R. F. Cavanagh and R. F. Waugh, “Secondary school renewal: The effect of classroom learning culture on educational outcomes,” Learning Environments Research , vol. 7, 2004, pp. 245–269.

[8] D. A. Freedman, Statistical models: Theory and practice, UK: Cambridge University Press, 2006.

[9] A. Lumpkin, “Caring teachers: The key to student learning,” Kappa Delta Pi Record, vol. 43, 2007, pp. 158-160.

[10] J. H. Stronge, Qualities of Effective Teachers, Alexandria, VA: Association for Supervision and Curriculum Development, 2007.

[11] K. Gould Lundy and L. Swartz, Creating Caring Classrooms: How to Encourage Students to Communicate, Create, and be Compassionate of Others. Markham, ON: Pembroke, 2011.

[12] J. Grzymala-Busse and W. Grzymala-Busse "An experimental comparison of three rough set approaches to missing attribute values". Transactions on Rough Sets , vol. 6, 2007, pp. 31–50.