Rough Sets as a Knowledge Discovery and Classification Tool for the Diagnosis of Students with Learning Disabilities

Tung-Kuang Wu
Dept. of Information Management, National Changhua University of Education
No. 2, ShiDa Rd., Changhua, Taiwan

Shian-Chang Huang*
Dept. of Business Administration, National Changhua University of Education
No. 2, ShiDa Rd., Changhua, Taiwan

Ying-Ru Meng
Dept. of Special Education, National HsinChu University of Education
No. 521, NanDa Rd., HsinChu City, Taiwan

Wen-Yau Liang
Dept. of Information Management, National Changhua University of Education
No. 2, ShiDa Rd., Changhua, Taiwan

Yu-Chi Lin
Dept. of Information Management, National Changhua University of Education
No. 2, ShiDa Rd., Changhua, Taiwan

Accepted: 02-12-2009
Received: 20-09-2010

Abstract

Due to the implicit characteristics of learning disabilities (LDs), the diagnosis of students with learning disabilities has long been a difficult issue. Artificial intelligence techniques like artificial neural network (ANN) and support vector machine (SVM) have been applied to the LD diagnosis problem with satisfactory outcomes. However, special education teachers or professionals tend to be skeptical to these kinds of black-box predictors. In this study, we adopt the rough set theory (RST), which can not only perform as a classifier, but may also produce meaningful explanations or rules, to the LD diagnosis application. Our experiments indicate that the RST approach is competitive as a tool for feature selection, and it performs better in term of prediction accuracy than other rule-based algorithms such as decision tree and ripper algorithms. We also propose to mix samples collected from sources with different LD diagnosis procedure and criteria. By pre-processing these mixed samples with simple and readily available clustering algorithms, we are able to improve the quality and support of rules generated by the RST. Overall, our study shows that the rough set approach, as a classification and knowledge discovery tool, may have great potential in playing an essential role in LD diagnosis.

Keywords: Rough Set, Knowledge Discovery, Learning Disabilities, LD Diagnosis.

1. Introduction

The term “Learning Disabilities” (LDs) was first used in 1963 [1]. However, experts in this field have not yet completely reach an agreement on the definition of LDs and its exact meaning [2]. According to definition given by the United States National Center for Learning Disabilities [3], a learning disability is:

“a neurological disorder that affects the brain's ability to receive, process, store, and respond to information.

* Corresponding author. No. 2, Shi-Da Rd, Changhua City, Taiwan. Tel. 886-937137456; Fax: 886-4-7211162, *Email:* shhuang@cc.ncue.edu.tw

The term learning disability is used to describe the seeming unexplained difficulty a person of at least average intelligence has in acquiring basic academic skills. These skills are essential for success at school and work, and for coping with life in general. LD is not a single disorder. It is a term that refers to a group of disorders."

As a result, a person can be of average or above average intelligence, without having any major sensory problems (like blindness or hearing impairment), and yet struggle to keep up with people of the same age in learning and regular functioning.

Due to the implicit characteristics of learning disabilities, the identification of students with LDs has long been a difficult and time-consuming process. In the United States, the so called "Discrepancy Model" [4], which states that a severe discrepancy between intellectual ability and academic achievement has to exist in one or more of these academic areas: (1) oral expression, (2) listening comprehension (3) written expression (4) basic reading skills (5) reading comprehension (6) mathematics calculation, is commonly adopted to evaluate if a student is eligible for special education services.

In Taiwan, the diagnosis procedure pretty much follows the "Discrepancy Model" and is roughly separated into 4 steps: (1) application for screening of potential students with LDs by parents, general education teachers and/or junior-level evaluation personnel, (2) identification of potential students with LDs by junior-level evaluation personnel, (3) diagnosis of possible students with LDs by senior-level evaluation personnel, and (4) final confirmation by special education specialists (usually college or university professors with LD major) [5]. Note, both junior-level and senior-level evaluation personnel are selected special education teachers with days' (junior level) or weeks' (senior level) training on LD diagnosis related procedure.

The sources of input parameters required in such prolonged process include information from parents, general education teachers, students' academic performance and a number of standard achievement and IQ tests. To guarantee collection of required information regarding to students suspected with LDs, usually checklists of some kind are developed to assist parents and general education teachers. The Learning Characteristics Checklists (LCC), a Taiwan locally developed LD screening checklist, is commonly used in some counties of Taiwan [6]. LCC consists of six features, which include LCC full scale index (LCC-FSI), LCC-A, LCC-B, LCC-C, LCC-D and LCC-E. Among the standard tests, the Wechsler Intelligence Scale for Children, Third Edition (WISC III) plays the most important role in the third and fourth stages of the current LD diagnosis model. The WISC-III is composed of 13 sub tests [7]. The scores of the sub-tests are then used to derive 3 IQs, which include full scale IQ (FIQ), verbal IQ (VIQ), performance IQ (PIQ), and 4 indexes, which include verbal comprehension index (VCI), perceptual organization index (POI), freedom from distractibility index (FDI), processing speed index (PSI) [7]. All IQ and index scores are normalized with a mean of 100 and a standard deviation of 15 [8]. There are also a number of locally developed standard achievement tests (AT), which typical consists of reading, math, and fields that related to students' academic achievement.

Diagnosis of students with LDs then involves mainly interpreting the standard tests scores and comparing them to the norms that are derived from statistical method. As an example, in case the difference between VIQ and PIQ is greater than 15, representing significant discrepancy between a student's cultural knowledge, verbal ability, etc, and his/her ability in recognizing familiar items, interpreting action as depicted by pictures, etc, is a strong indicator in differentiating between students with or without LD [6]. A number of similar indicators together with the students' academic records and descriptive data (if there is any) are then used as the basis for the final decision (by senior evaluation personnel and special education specialists). Confirmed possible LD students are then evaluated for one year before admitting to special education. However, it deserves to note that a previous study in Taiwan reveals that the certainty in predicting whether a student is having a LD using each one of the currently available indicators is in fact less than 50% [9].

As we can see, the above procedure involves extensive manpower (mainly the overloaded special education teachers) and resources. In addition, the diagnosis process requires that the special education teachers having a strong background in both psychology and statistics. Unfortunately, those were not commonly included in their training at the college level. Furthermore, a lack of nationally regulated standard for the LD diagnosis procedure and criteria results in

possible variations on the outcomes of diagnosis. In most cases, the difference can be quite significant [5]. Accordingly, the quality of interpretation varied and the pressure is primarily on the special education specialists at the final stage.

With the advance in artificial intelligence (AI) and its successful applications to various classification problems, it is interesting to investigate how these AI-based techniques perform in identifying students with LDs. In our previous study, we made attempts in adopting two well-known artificial intelligence techniques, artificial neural network (ANN) and support vector machine (SVM), together with various feature selection algorithms and evolutionary computation, to the LD diagnosis problem [5, 10]. The results are quite satisfactory, and indicate that AI may be a possible alternative solution to the problem. However, most special education teachers or professionals we talked to tend to be skeptical to this kind of black-box predictor. It is thus essential that we seek possible ways to combine our classifier with some other algorithms that can produce meaningful explanations or rules for the prediction. The thought leads us to the exploration of other potential technologies. Rough set theory is selected as it can be used for both feature reduction and classification, and at the same time generates rules that are meaningful to teachers and professionals in special education community.

The main objective of this study is thus to explore the feasibility of applying the rough set (RST) approach to the LD diagnosis problem. In the course of study, various pre-processing procedures, like clustering, feature discretization and reduction, will be applied to the collected data sets to evaluate their effect on the RST performance.

This rest of the paper is organized as follows. Section 2 briefly describes history of AI techniques on the special education applications, the rough set theory and its advantages over other classification methods. Section 3 and 4 presents the experiment settings, design and corresponding results. Finally, Section 5 gives a summary of the paper and lists some issues that deserve further investigation.

2. Related Work

Artificial intelligence techniques have long been applied to special education community. However, most attempts occurred in more than one or two decades ago and mainly focused on using the expert systems to assist special education in various ways [5].

In addition to expert systems, numerous machine learning based classification techniques have been developed and widely used in various applications [11]. Among all the classification techniques, artificial neural network (ANN) has received lots of attentions due to their demonstrated performance and has gained widely acceptance beginning from the 1990s [12]. The support vector machine (SVM) [13] has also emerged as a powerful tool for classification and performs better than artificial neural networks and other models in certain scenarios. A particular advantage of SVM over ANN is that it can be analyzed theoretically using concepts from computational learning theory, and at the same time can achieve good performance when applied to real world problems.

Our previous experiences in applying the above two classification approaches (ANN and SVM) to the LD diagnosis procedure have shown that ANN can achieve better performance in term of classification accuracy than the SVM model [5]. Unfortunately, both of the ANN and SVM techniques are among the so called black-box models and their generated results are difficult to interpret.

On the other hand, rough set theory (RST), proposed by Zdzislaw Pawlak in 1982 [14] to analyze the classification of uncertain or incomplete data, has a number of advantages over the above two models. Although both fuzzy set and rough set are used to deal with uncertain information, the RST is suitable for identifying relationships that might not be found using statistical methods [15]. The RST is a model of approximate reasoning, which can be used to manage vague and uncertain data or problems related to information systems, indiscernibility relations and classification, attribute dependence and approximation accuracy, reduct and core attribute sets, and decision rules [16].

The starting point of the RST is the assumption that some knowledge is associated with every object of interest. For example, if the object is a personal computer (PC), the PC's attributes may correspond to the data related to its functionalities. With the RST, real world information is represented by information table (*IT*). A row in the *IT* may describe an event, a patient, or an object. A column then represents an attribute of the event, patient, or object. Information table can be

defined as: $IT=(U, A)$, $U = \{x_1, x_2, \ldots, x_n\}$, representing a finite set of samples of size n, while $A = \{a_1, a_2, \ldots, a_p\}$, representing a finite set of attributes of size p. In most real world applications, these attributes may consist of p-1 conditional attributes and a decision attribute. In this case, the *IT* may be considered as a decision table (*DT*) [15].

Given a *DT*, it is possible that inconsistency, defined by objects with the same conditional attribute values yet have opposite consequences (decision), exists. In that case, approximation is used in RST to draw conclusion from the *DT*. In particular, the *lower approximation* of a set *X* depicts the set of objects with respect to *DT* that can be certainly classified as an equivalent class with the given conditional attributes. On the other hand, the *upper approximation* of a set *X* contains the set of objects that may possibly be classified as an equivalent class with the given conditional attributes. The difference of the upper and lower approximation of a set *X* is called the boundary region. Accordingly, certain rules may be drawn from the lower approximation of a set [15].

An original *DT* may contain redundant information, which includes *indiscernible* objects or superfluous attributes. To be more specific, objects are indiscernible if they are characterized by the same information. In RST, *indiscernibility* relation of objects means different objects with the same attribute values, which is the mathematical basis of rough set theory. Redundant information may be removed from the *DT* as long as it preserves data consistency, which leads to another essential idea of RST – the *reduct*. A *reduct* of a *DT* is a set *A* ($\subset DT$) that has the same indiscernibility information as the *DT* and the set *A* can not be further reduced. In other words, a reduct is a minimal sufficient subset of a set of attributes that has the same ability to discern concepts as when the full set of attributes is used [17]. They also represent necessary condition attributes in decision making. Accordingly, the RST can also be used to reduce data size and dimensionality in data analysis [18]. As a matter of fact, many researches have adopted the RST as a tool for feature selection [19].

To implement the rough set theory, a procedure, which includes generating reducts and identifying the decision rule, for determining the reducts is necessary. A number of algorithms and tools have been proposed and implemented to calculate the reducts associated with the RST [20, 21, 22]. To identify or compose the candidate reduct rules, a rule identification algorithm is developed based on [22], which includes four steps:

Step 1: Creating basic units and put into Database.
Step 2: Calculating the lower and upper approximations for basic units.
Step 3: Finding the core and reduct of attributes.
Step 4: Finding the core and reduct of attributive values.

In applications, the RST has lead to significant advances in many areas including knowledge discovery, machine learning, and expert systems [23]. For example, researchers proposed an approach to illustrate formulation of more meaningful rules using the notion of ordinal prediction. It proved to be an improvement for rule learning both in computing performance and in the usefulness of the rules derived from a case study on melanoma data [24]. Zhao *et al.* made an empirical experiment for letter recognition to demonstrate the usefulness of the discussed relations and reducts [25]. Tseng and Huang introduced a rough set theory application for feature selection in customer relationship management (CRM) [26]. Yang *et al.* presented a case study of applying rough set theory to analyze customer complaints in an IC packaging foundry in Taiwan [27]. One research discovers classification rules through a knowledge induction process that selects decision rules with a minimal set of features for real-valued data classification [28]. Jian *et al.* extended outlier detection to rough set theory, which has become a popular method for knowledge discovery in databases (KDD), much due to RST's ability to handle uncertain and/or incomplete data. Experimental results on real data sets also demonstrate the effectiveness of the RST method for outlier detection [29]. Yang and Wu applied rough sets to identify the set of significant symptoms causing diseases and to induce decision rules using the data from a Taiwan's otolaryngology clinic. Experimental results discover that the pattern is considered to be potentially helpful in improving the medical diagnosis [30].

The above examples show the versatility of the RST, which leads us to the thought that it should also have the potential in uncovering rules other than that are used in current LD diagnosis procedure or answering questions that may be currently under controversy in learning disabilities community.

In case the data attribute values processed by the RST are continuous, discretization of such real value

attributes is required prior to rules induction so as to reduce the number of rules while at the same time preserve the knowledge contents or the discernibility [31]. Many discretization algorithms have been proposed in the field of data mining [32].

In addition to the RST, clustering has also been applied to reduce data uncertainty due to outdated sources or imprecise measurement in order to achieve higher quality data mining results [33]. In general, clustering operates by organizing unlabelled data into groups of similar objects. Clustering in itself finds various applications in fields like marketing, bio-medical, web, and many others [34]. Among many proposed clustering algorithms, k-means [35] and two-step [36] are two commonly seen techniques that are available in various data mining tools.

In this paper, besides evaluating the potential of the RST in LD identification problem, it is also our objective to know whether applications of pre-processing procedures like clustering, discretization or reduct calculation help in improving the rules induced by the RST for the diagnosis of LD students.

3. Experimental Settings and Design

To fulfill the objectives of this study, we have designed and conducted four experiments to evaluate the potential of the RST approach as a knowledge discovery and classification tool for the identification of students with LDs. Combinations of various discretization and feature reduction algorithms will be explored to see how they perform on our collected data. Clustering will also be included at some point of the experiments to see how this uncertainty data reduction method affects the performance (e.g., quality of generated rules) of the RST. Finally, we also incorporate special education (or statistics) domain knowledge to determine more appropriate cut-points for data discretization. The results will be compared to those derived using other discretization algorithms.

The tools we use in this study include RSES [18], Rosetta [37], and YALE [38]. Five data sets contain test samples collected from counties located in the northern, central, and southern Taiwan (as shown in Table 1) are used as the training or validation data. Depending on the data sets, each sample may contain features from achievement test (AT), learning characteristics checklist (LCC), and/or WISC-III standard test.

Table 1. Data sets and their features used in this study

data set	sample size	percentage of students with LDs	feature×size†
A	125	19.5%	WISC-III×7, LCC×6, AT×3
B	159	47.8%	WISC-III×7, LCC×6, AT×3
C	656	25.0%	WISC-III×7, WISC-III×13
D	441	35.6%	WISC-III×7, WISC-III×13
E	878	54.2%	WISC-III×7

† AT represents achievement test, for dataset A it includes Word Recognition (WR), Reading and Math sub-tests, while for dataset B it includes Chinese, English and Math sub-tests. LCC represents learning characteristics checklist, which contains 6 features. Please refer to [6] regarding details of LCC. WISC-III×7 includes three IQ scores and 4 indexes, while WISC-III×13 includes the 13 WISC subtests. Please refer to [7] for further details on WISC-III standard test.

Among these five data sets, data set A, B, and C have been used extensively in our previous study [5, 10]. In particular, cases contained in data set A and C represent ones that follow a stricter diagnosis procedure as described in Section 1, with the one-year post evaluation executed by trained special education teachers [5]. On the other hand, although pretty much follow the same procedure, cases in data set B are diagnosed without involving special education specialists and with the one-year evaluation process conducted mostly by general education teachers. The somewhat looser procedure may have higher possibility in mistakenly diagnosing underachiever as having learning disabilities [5]. The latter two data sets, D and E, have just been acquired recently and included in this study. The source of data set D is the same as data set C, but with samples coming from later years. Data set E contains samples from central Taiwan, which is completely new to us. However, one thing deserves attention is that its percentage in diagnosing students as having LDs is much higher than the other data sets, which implies that its diagnosis criteria may be somewhat looser than the other counties.

The design of the four experiments are listed and explained in the following four sub-sections. Note that rules generated by RST are expressed in a form like: *If Conditions* (*C*) *then Decision* (*D*). The quality of such rules can be expressed by *certainty* and *coverage* factors, defined as follows [15].

$$Certainty = \frac{(\text{number of cases satisfying } C \text{ and } D)}{(\text{number of cases satisfying } C)} \quad (1)$$

$$Coverage = \frac{(\text{number of cases satisfying } C \text{ and } D)}{(\text{number of cases satisfying } D)} \quad (2)$$

Additionally, the number of cases that satisfy *C* is also referred to as *support*. For evaluation of ANN classification model, a performance index, correct identification rate (*CIR*), is defined as follows.

$$CIR = \frac{(\text{number of correct LD and non-LD identification})}{(\text{total number of cases})}$$

3.1. *Experiments 1*

The first experiment served as a preliminary study so that we can compare the RST performance to our earlier studies using ANN model. In addition, we would also like to see how the RST performs as a tool for feature reduction/selection, and how is the quality of rules generated by the RST. Data set A and B are chosen as the test samples since we have pretty much experience on both data sets and are very familiar with them.

Table 2. Procedure of Experiment 1

Repeat the following procedure twice with data set A and B being the *training* and *testing data-set* interchangeably
For *discretization-algorithm* = {global, local}
Perform *discretization-algorithm* on *training data-set* and output the *discretized-training-data-set* and *cut-off points* of each feature
Perform *discretization-algorithm* on *testing data-set* with *cut-off points* and output the *discretized-testing-data-set*
For *reduct-algorithm* = {exhaustive, genetic}
Perform *reduct-algorithm* on *discretized-training-data-set* and output the *reducts*
With each *reduct*, extract samples with associated attributes from *discretized-training-data-set* and output the *feature-reduced discretized-training-data-set*
Perform simple validation with the RST generated rules on the *feature-reduced discretized-training-data-set*
Output certainty / coverage factors
Select rules with higher certainty and support, validate each rule on the *discretized-testing-data-set*
Output each individual rule, its certainty and support factors

In this experiment, the input data set is first discretized, followed by a reduct generation process. For each selected feature set (reduct), a simple validation test with the input data set being randomly divided into two halves, each serves as the training (contain 60% of the samples) and validation (the rest 40% of the samples) data. In addition, rules with higher support and certainty in the above procedure are extracted and validated one by one on the other data set. The above procedure is repeated twice with roles of data set A and B interchanged. The procedure of experiment 1 is depicted in Table 2.

The tool we used in this experiment is RSES [18], which adopts Boolean reasoning approach to discretize data samples (referred to as local and global methods in RSES). For reducts and/or rules calculation, RSES use algorithms like exhaustive, genetic, dynamic, covering, and LEM2 algorithms. The later two methods are for rule generation only. Only the best results after trying all possible combinations of the above algorithms are output. Unless otherwise specified, RSES's default settings are used throughout the experiment.

3.2. *Experiments 2*

The objectives of this experiment are (1) to find possible combination(s) of discretization and reduct algorithms, and (2) to evaluate the three WISC-III feature sets (WISC-III×7, WISC-III×13, and WISC-III×20) that achieve better rule quality. The experiment proceeds by subsequently pre-processing the input data set with selected features by combinations of various discretization and feature reduction algorithms. A five-fold cross validation test is then performed on the pre-processed data set. We then measure the overall certainty and coverage by averaging the certainty and coverage of the tests. The procedure is depicted in Table 3.

In addition to RSES, Rosetta [37] is also used in this experiment so that we may be able to experiment with more discretization and reduct calculation algorithms. Note, Rosetta does include some RSES functionalities, but some of those may not be applicable to data samples larger than some predetermined size. In that cases, we use RSES instead. To differentiate between the two, algorithms derived from (or available in) RSES will be prefixed with "RSES" hereafter.

Prior to the cross validation test, the experiment starts by subsequent application of combinations of six discretization algorithms (RSESlocal, RSESglobal, entropy scaler, EFW scaler, naïve scaler, and semi-naïve

scalers) and five feature reduction algorithms (Johnson, Holte's, RSESexhaustive, RSESGenetic, and RSESdynamic reducers). For further information on the above mentioned data discretization and feature reduction algorithms, please refer to [18, 32] for more details.

Table 3. Procedure of Experiment 2

data-set = data set C
For *feature-set* = {WISC-III×7, WISC-III×13, and WISC-III×20}
For *discretization-algorithm* = {RSESlocal, RSESglobal, entropy, EFW, naïve, or semi-naïve scaler}
Perform *discretization-algorithm* on *data-set* containing *feature-set* and output the *discretized-data-set*
For *reduct-algorithm* = {Johnson, Holte's, RSESexhaustive, RSESdynamic, or RSESGenetic}
Perform *reduct-algorithm* on *discretized-data-set* and output the *reducts*
Perform five-fold cross-validation with the features listed in *reducts* on *discretized-data-set* for RST rules induction and validation
Output certainty / coverage factors and combinations of (*discretization-algorithm*, *reduct-algorithm*) that achieve the certainty / coverage

The samples we use in this experiment are from data set C. The reason for such a choice are twofold, (1) we have used data set C in our earlier study and thus are more familiar with this data set, and (2) data set C contains more samples than the others (e.g., data set A and B) so that we may have a more credible outcomes with this experiment. In addition, three features combinations (WISC-III×7, WISC-III×13, and WISC-III×20) of data set C are tested independently.

As a basis for comparison, we also include two well known rule generating algorithms, C4.5 and Ripper, in our study. C4.5 is an algorithm for the construction of a decision tree [39], while Ripper (Repeated Incremental Pruning to Produce Error Reduction) is a rule induction algorithm that was proposed by Cohen [40].

3.3. *Experiment 3*

In the third experiment, we try to use clustering to pre-process the data sets prior to the RST rules generation procedure. The training samples in this experiment are from data set A and B. For induced rules to be generalized, we retain only WISC-III×7 features that are common to all the five data sets. The objective of this experiment is to see whether the rules quality can be improved by excluding potential outliers contained in the data-sets with clustering. The procedure is depicted in Table 4.

Table 4. Procedure of Experiment 3

For *data-set* = {data set A, data set B, data set A∪B}
If clustering = YES
Perform k-mean and two-step clustering algorithms on *data-set* with number of cluster=2
Let *clustered data-set* = *data-set* ∩ {clustered cases that agree on both of the two clustering algorithms and experts' diagnosis}
Let (*discretization-algorithm*, *reduct-algorithm*) be the combinations that achieve higher certainty in Experiment 2
Perform *discretization-algorithm* on *data-set* / *clustered data-set* and output the *discretized-data-set*
Perform *reduct-algorithm* on *discretized-data-set* and output the *reducts*
Perform RST rules induction with the *reducts* and output the *generated-rules*
For *rule* in *generated-rules*
Validate the *rule* on data sets C, D, and E (after being discretized with *discretization-algorithm*) and output the certainty and support factors

Note that the procedure shown in Table 4 will be repeated three times, with data set A, B, or A∪B being processed, respectively. With each input data set, clustering step may or may not be applied to the samples before feeding them to the RST rule induction procedure. The clustering step is done by independently applying two clustering algorithms (K-means and two-step) to the data sets and then keeps only those samples that both of the two clustering algorithms and the experts' diagnosis all agree upon (by experts' diagnosis, we mean diagnosis that follows the procedure that we described in Section 1). Note the reason that we use K-means and two-step algorithms is because they happen to be available in the tool we used. Although it is not the focus of this study, we does conduct a simple experiment to evaluate how these two clustering algorithms, when applied individually or combined together, affect certainty and coverage of the RST induced rules.

The idea of combining data set A and B is coming from findings in our previous study [5]. To be more specific, we have noticed that ANN models generated from data set A is doing very well in predicting students

with learning disabilities. On the other hand, ANN models generated from data set B seem to perform better (as compared to those generated from data set A) in predicting students without learning disabilities. The difference may be resulted from inconsistency in the diagnosis process between these two counties [5]. Thus it seems intuitive to pre-process the combined data sets so that we may filter out samples that do not match in predictions by both clustering algorithms and the experts' diagnosis decision. It is expected that some falsely diagnosed cases can be excluded, and thus to potentially improve the overall quality of the RST generated rules.

Finally, the rules that generated from such a procedure are validated using data set C, D, and E, with those rules that have higher certainty being output. Note that in the discretization and feature reduction procedures, only combinations of the two algorithms producing better predictions in experiment 2 are included.

3.4. *Experiments 4*

The objective of experiment 4 is to compare the results (in terms of rules quality and support) of using manual discretization and discretization algorithm(s) that performed better in experiment 3. Accordingly, the experiment is pretty much the same as experiment 3, except that the discretization procedure is done manually according to the fact that both WISC-III IQs and indexes have been normalized to a mean of 100 and a standard deviation of 15 [8]. All IQ and index scores are then discretized into six intervals with 100 being the center cut-point and all other cut-points set to 100 plus/minus one or two times the standard deviation. Accordingly, the six intervals include [*, 70), [70, 85), [85, 100), [100, 115), [115, 130) and [130, *). The notation [*x*, *y*) represents the range of score is greater than or equal to value *x* and less than value *y*. In addition, [*, *y*) or [*x*, *) indicates that the interval is less than *y* or greater than or equal to *x*, respectively. In addition, only the data set (or combined data sets) that produced the best prediction rules in experiment 3 is used. The procedure is depicted in Table 5.

Table 5. Procedure of Experiment 4

Let *data-set* = data-set in {data set A, data set B, data set A ∪ B} that performs best in experiment 3
If clustering = YES
Perform k-mean and two-step clustering algorithms on *data-set* with number of cluster=2
Let *clustered data-set* = *data-set* ∩ {clustered cases that agree on both of the two clustering algorithms and experts' diagnosis}
Let *reduct-algorithm(s)* = algorithm(s) that achieve the highest certainty in Experiment 2
Perform *manual discretization procedure* (with 70, 85, 100, 115 and 130 being the cut-points) on *data-set / clustered data-set* and output the *discretized-data-set*
Perform *reduct-algorithm* on *discretized-data-set* and output the *reducts*
Perform RST rules induction with the *reducts* and output the *generated-rules*
For *rule* in *generated-rules*
Validate the *rule* on data sets C, D, and E (after being discretized with *manual discretization procedure*) and output the certainty and support factors

4. Results and Implications

In the following, we will present results of the four experiments depicted in Section 3, together with our findings and interpretations.

4.1. *Results of Experiment 1*

By applying the two data reduction algorithms that RSES provides (exhaustive and genetic algorithms) to the discretized data set (using global or local discretization methods), we have calculated the corresponding reducts for data set A and B. For the purpose of comparison, a simple validation procedure using ANN model, as depicted in [5, 10], is also preformed on the data sets with selected features (reduct sets). The selected feature sets, simple validation test results using the RST and ANN model of data set A and B are shown in Table 6 and 7, respectively. Note, in each iteration of experiment 1, only four selected reducts with higher simple validation certainty (using the RST) are shown (one by global discretization and three by local discretization).

Based on our previous experiences with other feature selection algorithms (wrapper-based GA algorithm) [5, 10], we are impressed by the quick processing time, which is in a matter of seconds, of the RST approach in generating these feature sets. As a comparison, for the same data set, depending on the wrapped learner, a GA-based feature selection procedure usually take tens of minutes or even hours for producing just one feature set. However, the questions remain would be: (1) as a classification tool, is the RST

better than the ANN model? (2) as a feature selection tool, is the quality of selected features by the RST comparable to the other approaches?

Table 6. Selected feature sets and their corresponding best certainty/CIR with data set A. (The number within parenthesis represents the coverage of the prediction.)

	discret. alg.	selected features	Certainty with RST	CIR with ANN
1	global	WR, Math, LCC-B, LCC-D, LCC-E, VCI, POI, PSI	86% (1.0)	86% (1.0)
2	local	WR, Reading, Math, LCC-A, LCC-D, PIQ, VCI, POI, PSI	86% (1.0)	**90%** (1.0)
3	local	WR, Reading, LCC-D, LCC-E, PIQ, VCI, POI, PSI	82% (1.0)	82% (1.0)
4	local	WR, Reading, Math, LCC-A, LCC-D, LCC-E, PIQ, VCI, POI	74% (1.0)	86% (1.0)

Table 7. Selected feature sets and their corresponding best certainty/CIR with data set B. (The number within parenthesis represents the coverage of the prediction.)

	discret. alg.	selected features	Certainty with RST	CIR with ANN
1	global	Chinese, English, LCC-A, LCC-B, LCC-C, LCC-E, VIQ, FDI, PSI	76.6% (1.0)	85.9% (1.0)
2	local	Chinese, LCC-B, LCC-C, VIQ, PIQ, FIQ, POI, FDI, PSI	84.4% (1.0)	90.6% (1.0)
3	local	Chinese, Math, LCC-B, LCC-C, VIQ, FIQ, FDI, PSI	82.8% (1.0)	85.9% (1.0)
4	local	Chinese, Math, LCC-B, LCC-C, VIQ, PIQ, FIQ, FDI	81.2% (1.0)	85.9% (1.0)

For the first question, according to results shown in Table 6 and 7, the RST is a little bit behind in most cases as a classification tool for identifying students with LDs. However, as a tool for feature selection, the RST seems to be competitive to the wrapper-based genetic approach in a number of cases. In particular, the CIR using the second feature set with ANN model in Table 6 is 90%, which is the highest that we have ever got from data set A (in term of simple validation). As a comparison, 86% is the best we achieved in our earlier studies by ANN model [5, 10]. On the other hand, although the best feature set selected from data set B using the RST may achieve 90.6% in CIR (see Table 7) by ANN model, yet it may still fall a little short from what we achieved in [5] (which would be 93.8% in CIR).

Up to this point, we already know how the two classifiers (the RST and ANN) perform in term of LD identification accuracy. Usually, for black-box predictor like ANN, this would be the end of discussion. However, a favorable characteristic of RST is that it can not only produce a model based on existing data so as to classify new cases, but it also provides us the opportunity to analyze the model and gain new insight into the problem [32]. As a further illustration, the RST approach may be able to generate rules like the ones listed in Table 8 and 9. The fact that the "classification model" being represented in a form (i.e., rules as shown in Table 8 and 9) familiar to specialists from the special education community does make the RST look more appealing than the other approaches.

Table 8. Rules extracted from experiment 1 using data set A, with each rule being the one that receives the top-four most support within its class. (The number within parenthesis represents the support when applying the rule to the specific data set.)

	Rules	certainty in data set A	certainty in data set B
1	(LCC-B > 84.5) & (VCI < 87) & (POI < 97) → LD=**NO**	100% (46)	75.5% (49)
2	(POI < 97) & (PSI < 92) & (VCI < 87) → LD=**NO**	100% (44)	77.9% (77)
3	(LCC-D < 83) & (VCI < 87) & (POI < 97) → LD=**NO**	100% (41)	74.6% (59)
4	(VCI < 87) & (POI < 97) → LD=**NO**	98.3% (58)	72.7% (99)
5	(Math < 30.5) & (POI > 97) & (PSI < 92) → LD=**YES**	100% (11)	—†
6	(LCC-D < 83) & (POI > 97) & (PSI < 92) → LD=**YES**	100% (10)	100% (12)
7	(VCI < 87) & (POI > 97) & (PSI < 92) → LD=**YES**	100% (8)	100% (10)
8	(LCC-B > 84.5) & (POI > 97) & (PSI < 92) → LD=**YES**	100% (9)	100% (8)

† The rule can not be generalized to data set B as Math scores between the two data sets are not standardized.

Table 9. Rules extracted from experiment 1 using data set B, with each rule being the one that receives the most support within its class. (The number within parenthesis represents the support when applying the rule to the specific data set.)

	Rules	certainty in data set B	certainty in data set A
1	(VIQ < 68) & (FDI < 71) → LD=**NO**	100% (41)	87.5% (88)
2	(PIQ < 80) & (POI < 72) → LD=**NO**	100% (36)	84.6% (26)
3	(VIQ < 68) & (POI < 72) → LD=**NO**	100% (34)	100% (11)
4	(POI < 72) & (FDI < 71) → LD=**NO**	100% (26)	100% (7)
5	(Chinese < 15) & (LCC-C > 85) & (VIQ > 74) & (FDI < 90) → LD=**YES**	100% (31)	—†
6	(FIQ > 88) & (70 < FDI < 87) → LD=**YES**	100% (27)	50% (10)
7	(LCC-C > 89) & (FIQ > 88) → LD=**YES**	100% (21)	33.3% (12)
8	(LCC-C > 89) & (VIQ > 74) & (PSI < 83) → LD=**YES**	100% (15)	11.8% (17)

† The rule can not be generalized to data set A, which does not contain Chinese feature.

After reviewing rules in Table 8 and 9, one may notice an interesting phenomenon. It appears that "YES" rules induced from data set A can be generalized quite well to the other data set. For example, the three "YES" rules (rule #6~8 of Table 8) are able to correctly identifies students with LDs in data set B without any false positive. This is quite a remarkable performance if they can be validated with further research and generalized to more samples. The same goes to the "NO" rules generated from data set B. Among the four "NO" rules listed in Table 9, two of them (rule # 3 and 4) can also filter non-LD samples from data set A without any false negative. While the other two are having certainty around 85%. The implication behind is that the burden of special education teachers or evaluation personnel can be somewhat relieved as they have fewer cases to evaluate. In addition, due to the effect of features reduction, they do not have to take into account a large number of features, either. On the other hand, the "NO" rules (or the "YES" rules) induced from data set A (or data set B) do not seem to generalize equally well. In some cases (e.g., rule #6~8 of Table 9), the certainty factors of applying such rules to the other data set are less than 50%.

The above outcomes once again confirm our earlier findings regarding inconsistency in the diagnosis process between the two counties that we acquired data set A and B [5]. In addition, according to the fact that "YES" rules from data set A can be generalized with strong certainty indicates that the county from which we collected data set A indeed follows a stricter diagnosis procedure.

When look closer into the rules with higher certainty, we find some common sub-rules occurred repeatedly. For examples, (POI > 97) & (PSI < 92) within "YES" rules. The sub-rule appears to conform to earlier study stating relatively that students with LDs usually have their PSI score lower than POI score [41]. Apparently, the results by the RST go one step further by indicating the absolute values of the two indexes. It is thus our belief that by cross examining findings from the special education community and rules induced by the RST carefully, we may be able to uncover step by step more useful information behind the LD diagnosis problem.

4.2. *Results of Experiment 2*

The results of experiment 2 are shown in Table 10. Note that only results with certainty factor above 90% are listed. A number of observations can be derived according to the data presented.

First of all, the WISC-III×7 feature set, containing three IQs and four indexes, appears to be the best features combination. The WISC-III×20 feature set comes in second. On the other hand, the thirteen sub-test scores of WISC-III (WISC-III×13) do not seem to have much effect in the RST rule induction procedure.

Second, discretization using RSESlocal, RSESglobal, and naïve scalers seem to have the best positive effect to the improvement of certainty. When considering the feature reduction algorithms, RSESexhaustive, RSESgenetic, RSESdynamic, and Johnson algorithms seem to perform equally well.

Finally, the RST approach seems to perform better than C4.5 and Ripper in term of certainty if not taking into account the coverage rate. It even performs better than results from our earlier study by combining evolutionary computation with ANN model (which would be 86.7% in CIR [10]). Note, we need to point out that from the point-of-view of special education community, practitioners may be equally or even more concerned with higher precision in positive identification of students with LDs (or filtering out

students without LD), even though with lower coverage rate. In addition, according to the results, both of C4.5 and Ripper algorithms may seem to be benefited from the pre-processing steps with applications of appropriate discretization and feature reduction algorithms.

Table 10. Five-fold cross validation results of experiment 2 on data set C using rough set, C4.5, and Ripper algorithms. Only combination(s) of discretization and reduct algorithms that achieve better rule quality (by RST) in terms of certainty and coverage (listed in parentheses) are shown.

Feture set	discret. alg.	Reduct alg.	RST (%)	C45(%) (w/wo†)	Rip (%) (w/wo†)
WISC-III×7	RSES Local	RSES exhaustive/ genetic	100 (.19)	80/81 (1)	80/82 (1)
WISC-III×7	RSES Global	RSES exhaustive/ genetic	100 (.16)	80/81 (1)	80/82 (1)
WISC-III×20	Naïve	Johnson	93 (.09)	81/81 (1)	82/82 (1)
WISC-III×7	Naïve	RSES dynnmic	92 (.33)	80/81 (1)	80/82 (1)
WISC-III×7	Naïve	RSES exhaustive/ genetic	92 (.33)	80/83 (1)	80/83 (1)
WISC-III×7	Naïve	Johnson	90 (.07)	80/80 (1)	80/79 (1)

† In cases of C4.5 and Ripper, certainty with (w) or without (wo) discretization and reduction pre-processing are shown.

4.3. *Results of Experiment 3*

With the knowledge derived form experiment 2, we retain only RSESlocal / RSESglobal / naïve scaler and RSESdynamic / RSESexhaustive / RSESgenetic / Johnson feature reduction algorithms in experiment 3. For each input data set, we choose to output at most six rules that result in the best certainty, both in identifying LD and non-LD students. The results are shown in Table 11 (with un-clustered input samples) and 12 (with clustered input samples). Note that rules generated from data set A or B alone are not shown since they perform no better than ones resulted from combining data set A and B. Accordingly, the term "pre-processing with clustering" (or similar) means specifically the scenario in which we combine the two data sets first and followed by application of clustering procedure.

Table 11. Rules generated from data set A∪B without clustering prior to rules induction.

	rules	certainty	support
1	(PIQ < 77) & (FIQ < 65) → LD=**NO**	100%	182
2	(PIQ < 74) & (FIQ < 76) & (VCI < 70) & (POI<72) → LD=**NO**	100%	130
3	(PIQ < 74) & (FIQ < 76) & (VCI < 70) & (FDI<71) → LD=**NO**	100%	115
4	(FIQ < 76) & (VCI < 70) & (POI < 72) & (FDI < 71) → LD=**NO**	100%	90
5	(PIQ < 74) & (VCI < 70) & (POI < 72) & (FDI < 71) → LD=**NO**	100%	89
6	(PIQ > 76) & (FIQ > 82) & (84 < VCI < 94) & (95 < POI < 102) & (70 < FDI < 76) & (86 < PSI < 99) → LD=YES	100%	5
7	(98 ≤ PIQ < 101) & (91 ≤ FIQ < 92) & (71 ≤ FDI < 112) & (PSI < 89) → LD=**YES**	80.0%	4
8	(84 ≤ FIQ < 91) & (VCI < 87) & (71 ≤ FDI < 112) & (89 ≤ PSI < 91)) → LD=**YES**	62.5%	10
9	(59 ≤ PIQ < 98) & (72 ≤ FIQ < 79) & (55 ≤ POI < 102) & (VCI < 87) & (68 ≤ FDI < 71) → LD=**YES**	62.5%	15
10	(72 ≤ FIQ < 79) & (VCI < 87) & (55 ≤ POI < 102) & (68 ≤ FDI < 71) → LD=**YES**	62.5%	15

By carefully reviewing these rules, we have the following findings. First, by mixing the training samples from different data sets (derived from counties that may have inconsistency in their diagnosis process) and pre-processing with clustering prior to rules induction, both certainty and support of the generated rules can be improved significantly. In other words, the clustering procedure that was conducted in the experiment may have effectively removed some data inconsistency, which again contributes to the quality of induction rules (in terms of certainty and support) by the RST. In particular, for positive LD diagnosis prediction, rules that generated with clustering step (Table 12) are having certainty factor above or closer to 90%, much higher than their counterpart (Table 11), which in most cases are just slightly higher than 60%. In case of support,

rules generated with clustering are higher in number in most cases. In addition, rules generated with clustering pre-processing usually involve fewer features and are thus simpler and more generalized. For example, rules in Table 11 contain slightly more than four features in average, while it is about three in Table 12.

Table 12. Rules generated from data set A∪B with clustering prior to rules induction.

	rules	certainty	support
1	(PIQ < 74) & (FIQ < 76) & (VCI < 70) → LD=**NO**	100%	171
2	(PIQ < 74) & (VCI < 70) & (POI < 72) → LD=**NO**	100%	130
3	(PIQ < 74) & (VCI < 70) & (FDI < 71) → LD=**NO**	100%	115
4	(VIQ < 64) & (PIQ < 74) & (VCI < 70) → LD=**NO**	100%	97
5	(VCI < 73) & (POI < 72) → LD=**NO**	98.8%	173
6	(VIQ < 95) & (FIQ < 87) & (PIQ < 91) & (VCI < 73) & (POI < 72) → LD=NO	98.8%	172
7	(106 ≤ PIQ < 110) & (VCI < 87) & (PSI < 89) → LD=**YES**	90.9%	11
8	(POI > 101) & (73 < FDI < 78) → LD = YES	90.0%	50
9	(FIQ < 87) & (POI > 100) → LD = YES	89.3%	75
10	(106 ≤ PIQ < 110) & (VCI < 87) & (71 ≤ FDI < 112) & (PSI < 89)) → LD=**YES**	88.9%	9
11	(98 ≤ PIQ < 101) & (105 ≤ POI < 108) & (71 ≤ FDI < 112) & (PSI < 89) → LD=**YES**	88.2%	17
12	(105 ≤ POI < 108) & (PSI < 89) → LD=**YES**	87.2%	39

Second, for non-LD prediction, rules generated by both procedures can all achieve 100% (or closer to 100%) in term of certainty. Their supported cases are also much higher than the LD prediction rules, which, as we have noted earlier, may effectively reduce the loading of special education teachers or evaluation personnel since they may have fewer cases to evaluate.

However, there are a couple of issues with the LD diagnosis prediction rules. First, we notice that some rules have quite a narrow margin, e.g., PIQ in rules 7, 10, and 11 or POI in rules 11 and 12 of table 12, which may pose a strict burden to the WISC-III test procedure and interpretation. Second, the support for the generated rules are much lower when compared to the total number of samples tested. For RST to be an essential part in the LD diagnosis related problem in the future, these two issues need to be addressed and resolved.

Finally, according to our observation, rules for non-LD prediction shown in Table 11 and 12 are quite similar to those generated from (the original or clustered) data set B only. On the other hand, rules for LD prediction are also similar to those derived from (the original or clustered) data set A alone. The observations also conform to what we derived in experiment 1 and our earlier study [5].

Note, in order to gain more insight into the effects of the clustering procedures on the RST induced rules, we conducted an additional simple sensitivity experiment by repeating part of experiment 3. We retain only data set A∪B, and then modify the subsequent clustering and validation steps. For clustering step, two more scenarios, which consist of applying only one of the two clustering algorithms (K-means or two-step) to the data set and keep those samples that match the clustering outcomes and the experts' diagnosis, are included. Accordingly, we now have un-clustered data set, K-means or two-step clustered data set, and K-means + two-step clustered data set. For validation step, instead of examining every individual rule, we select rules that have higher support (greater or equal to 10) and verify these rules on the other data sets to get the certainty and coverage. The corresponding results are presented in Table 13.

Table 13. A simple sensitivity study to evaluate the effect of the clustering procedure.

data set	certainty (coverage)			
	un-clustered	K-means only	two-step only	K-means+ two-step
C	89.4% (0.418)	51.6% (0.996)	49.2% (0.998)	**95.3% (0.454)**
D	87.6% (0.383)	61.7% (0.995)	58.7% (0.993)	87.5% (**0.454**)
E	56.4% (0.339)	59.4% (0.993)	58.9% (0.993)	**58.4% (0.411)**

According to the results, it is obvious that application of only one clustering algorithm is not enough to filter the potential outliers. On the other hand, we may see very clearly the improvement in terms of

both certainty and coverage by the K-means + two-step clustering procedure.

4.4. *Results of Experiment 4*

Similar to experiment 3, in this experiment, only the top five rules that result in the best certainty, both in identifying LD and non-LD students, are shown (refer to Table 14).

Table 14. Rules generated from data set A∪B with manual discretization and clustering prior to rules induction.

	Rules	Support	Certainty
1	(PIQ < 70) & (FIQ < 70) & (VCI < 70) => LD=**NO**	130	100%
2	(VIQ < 70) & (POI < 70) → LD=**NO**	167	100%
3	(VIQ < 70) & (PIQ < 70) & (VCI < 70) => LD=**NO**	127	100%
4	(PIQ < 70) & (VCI < 70) & (POI < 70) => LD=**NO**	118	100%
5	(PIQ < 70) & (VCI < 70) & (FDI < 70) => LD=**NO**	93	100%
6	(100 ≤ POI < 115) & (PSI < 70) → LD=**YES**	11	90.9%
7	(70 < VCI < 85) & (100 < POI < 115) & (85 < PSI < 100) → LD = YES	54	85.2%
8	(70 ≤ VIQ < 85) & (100 ≤ PIQ < 115) → LD=**YES**	89	84.0%
9	(100 ≤ PIQ < 115) & (70 ≤ VCI < 85) → LD=**YES**	69	84.0%
10	(70 ≤ VCI < 85) & (100 ≤ POI < 115) → LD=**YES**	107	82.0%

Overall, it appears that for LD prediction rules, the certainty is somewhat lower than those listed in Table 12, yet the rules are more concise and the issue with too few supports has been resolved. On the other hand, for non-LD prediction rules, the number of support is a bit lower than those shown in Table 12.

Upon reviewing the LD prediction rules more carefully, our colleague in special education acknowledges that rule #6 is not currently used in the LD diagnosis process and may deserve further investigation. On the other hand, rules #7~10 seem to fit into a well-known predictor, $|PIQ - VIQ| \geqq 15$ (note that VCI and POI are potential substitutes for VIQ and PIQ in some cases [7]), as stated earlier in Section 1. However, rules 6~8 are still valuable as they present not just the relative difference between the two IQs / indexes, but their absolute values, too.

To be more specific, a student with VIQ (or VCI) score one or two standard deviations below the average (i.e., between 70 and 85) has long been a difficult case for diagnosis. The major reason is that underachievers, students with mild mental retardation or learning disabilities may all have their VIQ (or VCI) score falls into this interval. Accordingly, it is very likely that a student with LDs may be misdiagnosed as an underachiever or one with mild mental retardation (or vice versa). Since the instructional objectives for students of these three categories are quite different (i.e., cognitive, functional or response-to-intervention for students with LDs, mild mental retardation and underachievers), the negative impact for such a misdiagnosis on the students can be enormous. Fortunately, rules #7~12 in Table 8 indicate that in case the PIQ (or POI) score falls between 100 and 115, it is most likely (with more than 80% certainty) that the case under consideration would be one with LDs. Accordingly, the information may potentially reduce the risk of misdiagnosis.

As IQ score of 70 being the decision boundary differentiating students with learning disabilities and metal retardation, non-LD prediction rules in Table 8 do not seem to present much surprise to special education practitioners. In comparison, non-LD prediction rules in Table 12 are more valuable since they indicate some more appropriate cut-points for filtering students with mental retardation from the LD diagnosis procedure. The outcome also implies that a discretization process incorporating too much special education (or statistical) knowledge might just reproduces rules that have already been known.

5. Conclusion and Future Research

The identification and diagnosis of students with learning disabilities, which requires a lot of man power, resources and expertise, have never been an easy job. Although ANN and SVM models have been applied to the LD diagnosis problem with satisfactory outcomes, special education teachers or professionals seem to be skeptical to these kinds of black-box predictors. Accordingly, in this study, we made an attempt to apply the RST approach, which can be used as a tool for classification, knowledge discovery, and most important of all, generating rules that are represented in a form

familiar and acceptable to practitioners in the special education community.

The preliminary results show that rough set classifier may not have the full coverage of samples like the other models, e.g., C4.5, Ripper, or ANN in our earlier study. However, RST approach does show its capability in discovering currently unknown knowledge behind the LD diagnosis procedure, which certainly helps special education specialists in finding new decision criteria for LD diagnosis. In particular, to the best of our knowledge, some of the rules discovered in this study have never been used or appeared in any LD diagnosis context. In addition, conventional rules derived from statistical method for the diagnosis of students with LDs usually involve only relative differences between various IQ or index scores. On the other hand, the RST generated rules specify some definite intervals with much higher diagnosis certainty, which would certainly be more useful to the LD diagnosis personnel. For example, rule #6~8 of Table 8 or rule #7~12 can all correctly identify LD students with no or around 10% false positives. This is quite encouraging to us since none of the currently available LD diagnosis criteria can achieve such a high degree of certainty [9]. Based on the observations described above, the primary contribution of this study is thus in demonstrating RST's potential in the LD diagnosis application.

The second contribution of this study would be the idea of incorporating clustering procedure to the mixed input samples prior to RST rules induction. The application of the clustering step on the mixed data sets collected from different sources is able to remove uncertain cases, which is especially essential in Taiwan as various counties may have quite different LD diagnosis procedure and criteria. The outcomes as a result of clustering are rules with better support and improved certainty.

In the future, we will work closely with our special education colleagues to verify the rules that are discovered in this study. In addition, we also noticed, from comparing results in Table 12 and 14, the discretization procedure can be an essential process affecting quality and support of the generated rules. In future study, we shall be working on integrating (carefully) special education domain knowledge with the existed discretization algorithms so that we can determine some appropriate cut-points that may uncover more precious and hidden information to assist the LD diagnosis procedure without reproducing rules that might have already been known. Finally, it may worth trying to adjust the parameter settings of the two clustering algorithms used in this study or adopting other clustering methods [42, 43] in the future to see if those make any difference to the rules generated.

Acknowledgements

This work was supported in part by the National Science Council, Taiwan, under Grant NSC 98-2511-S-018 -014 -MY2

References

1. S. A. Kirk, Behavioral Diagnosis and Remediation of Learning Disabilities. *Proc. of Conf. on the Exploration into the Problems of the Perceptually Handicapped Child*, (Evanston, IL, 1963), pp. 1–7.
2. J. M. Fletcher, W. A. Coulter, D. J. Reschly and S. Vaughn, Alternative Approach to the Definition and Identification of Learning Disabilities: Some Questions and Answers, *Annals of Dyslexia*, **54**(2) (2004), 304–331.
3. National Center for Learning Disabilities, http://ncld.org/ld-basics/ld-explained/basic-facts/what-are-learning-disabilities, extracted on Nov. 16th, 2009.
4. J. Schrag, *Discrepancy Approaches for Identifying Learning Disabilities* (National Association of State Directors of Special Education Alexandria, VA, 2000).
5. T.-K. Wu, S.-C. Huang and Y.-R. Meng, Evaluation of ANN and SVM Classifiers as Predictors to the Diagnosis of Students with Learning Disabilities, *Expert Systems with Applications*, **34**(3) (2008), 1846–1856.
6. Y.-R. Meng and L.-R. Chen, "On Discussing the Differences about the Learning Characteristics of LD," *Bulletin of Special Education*, **23** (2002), 75–93. (in Chinese).
7. C. L. Nicholson and C. L. Alcorn, Interpretation of the WISC-III and Its Subtests, *Proc. 25th Annual Meeting of the National Association of School Psychologists*, (Washington DC, 1993).
8. J. R. Reddon, S. V. Veen and J. E. Reddon, Seemingly anomalous full scale IQ scores on the WAIS-III and the WISC-III, *Current Psychology*, Springer, New York, **23**(1) (2004), 86–94.
9. T.-S. Huang, A Study on the Characteristics of WISC-III for Students With Learning Disabilities, Master thesis, Graduate Institute of Special Education, National HsinChu University of Education (in Chinese), 2006.
10. T.-K. Wu, S.-C. Huang and Y.-R. Meng, Improving ANN Classification Accuracy for the Identification of Students with LDs through Evolutionary Computation, *Proc. of the 2007 IEEE Congress on Evolutionary Computation* (Singapore, 2007).

`11. B. Baesens, T. Van Gestel, S. Viaene, M. Stepanova, J. Suykens and J. Vanthienen, Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring, *Journal of the Operational Research Society*, **54**(6) (2003), 627–635.
12. C. M. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, Oxford, UK, 1995).
13. C. Cortes, and V. Vapnik, Support-Vector Networks, *Machine Learning*, **20** (1995), 273–297.
14. Z. Pawlak, Rough sets. *Inl. Journal of Information and Computer Science*, **11**(5) (1982), 341–356.
15. Z. Pawlak, A Primer on Rough Sets: a New Approach to Drawing Conclusions from Data, *Cardozo Law Review*, **22** (2001), 1407–1415.
16. J. Y. Shyng, F. K. Wang, G. H. Tzeng and K. S. Wu, Rough Set Theory in Analyzing the Attributes of Combination Values for the Insurance Market, *Expert Systems with Applications*, **32**(1) (2007), 56–64.
17. W. Ziarko, Decision Making with Probabilistic Decision Tables, *Lecture Notes In Computer Science*, **1711** (1999), 463–471.
18. A. An, Y. Huang, X. Huang and N. Cercone, Feature Selection with Rough Sets for Web Page Classification, *Lecture Notes in Computer Science*, **3135** (2005), 1–13.
19. K. Thangavela and A. Pethalakshmib, Dimensionality reduction based on rough set theory: A review, *Applied Soft Computing*, **9**(1) (2009), 1–12.
20. J. Bazan and M. Szczuka, RSES and RSESlib - A Collection of Tools for Rough Set Computations, *Lecture Notes in Artificial Intelligence*, **2005** (2001), 106–113.
21. Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data* (Kluwer Academic Publishers, Boston, US, 1991).
22. C. C. Huang and T. Tseng, Rough Set Approach to Case-Based Reasoning Application, *Expert Systems with Applications*, **26**(3) (2004), 369–385.
23. Z. Suraj, An Introduction to Rough Set Theory and Its Applications–A Tutorial, *Proc. 1st Int. Computer Engineering Conference: New Technologies for the Information Society* (Cairo, Egypt, 2004).
24. P. Pattaraintakorn, N. Cercone and K. Naruedomkul, Rule learning: Ordinal Prediction Based on Rough Sets and Soft-Computing, *Applied Mathematics Letters*, **19**(12) (2006), 1300–1307.
25. Y. Zhao, Y. Yao and F. Luo, Data Analysis Based on Discernibility and Indiscernibility, *Information Sciences*, **177**(22) (2007), 4959–4976.
26. T. L. Tseng, C. C. Huang, Rough Set-based Approach to Feature Selection in Customer Relationship Management, *Omega*, **35**(4) (2007), 365–383.
27. H. H. Yang, T. C. Liu and Y. T. Lin, Applying Rough Sets to Prevent Customer Complaints for IC Packaging Foundry, *Expert Systems with Applications*, **32**(1) (2007), 151–156.
28. Y. Leung, M. M. Fischer, W. Z. Wu and J. S. Mi, A Rough Set Approach for the Discovery of Classification Rules in Interval-valued Information Systems, *International Journal of Approximate Reasoning*, **47**(2) **(**2008), 233–246.
29. F. Jiang, Y. Sui and C. Cao, Some Issues about Outlier Detection in Rough Set Theory, *Expert Systems with Applications*, **36**(3) (2009), 4680–4687.
30. H. H. Yang and C. L. Wu, Rough sets to help medical diagnosis–Evidence from a Taiwan's clinic, *Expert Systems with Applications*, **36**(5) (2009), 9293–9298.
31. C.-Y. Chen, Z.-G. Li, S.-Y. Qiao and S.-P. Wen, Study on Discretization in Rough Set Based on Genetic Algorithm, *Proc. of the Second Int. Conf. on Machine Learning and Cybernetics* (Xi'an, China, 2003).
32. S. H. Nguyen and H. S. Nguyen, Discretization Problems for Rough Set Methods. *Lecture Notes in Artificial Intelligence*, **1428** (1998), 545–552.
33. M. Chau, R. Cheng, B. Kao and J. Ng, Uncertain Data Mining: An Example in Clustering Location Data, *Lecture Notes in Artificial Intelligence*, **3918** (2006), 199–204.
34. P. Berkhin, Survey of Clustering Data Mining Techniques, http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf, extracted on Apr. 10th, 2010.
35. J. A. Hartigan and M. A. Wong, A K-means Clustering Algorithm, *Applied Statistics*, **28** (1979), 100–108.
36. M. Kayard, Two-Step Clustering Analysis in Researches: A Case Study, *Eurasian J. of Educational Research*, **28** (2007), 89–99.
37. A. Øhrn and T. Rowland, Rough Sets: A Knowledge Discovery Technique for Multifactorial Medical Outcomes, *American J. of Physical Medicine & Rehabilitation*, **79**(1) (2000), 100–108.
38. Rittho, O., R. Klinkenberg, S. Fischer, I. Mierswa and S. Felske, YALE: Yet Another Learning Environment, *LLWA 01-Tagungsband der GI-Workshop-Woche Lernen-Lehren-Wissen-Adaptivität* (Dortmund, Germany, 2001).
39. J. R. Quinlan, *C4.5: Programs for Machine Learning* (Morgan Kaufmann Publishers, 1993).
40. W. Cohen, Fast Effective Rule Induction, *Proc. of Int. Conf. on Machine Learning* (Lake Tahoe, CA, 1995).
41 H.-Y. Chen and T.-R. Yang, Base Rates of WISC-III Diagnostic Subtest Patterns in Taiwan: Standardization, Learning Disabled and ADHD Samples Applied, Psychological Testing, **47**(2) (2000), 91–110. (in Chinese).
42 R. K. Brouwer, Fuzzy Relational Fixed Point Clustering, *International Journal of Computational Intelligence Systems*, **2**(1) (2009), 69–82.
43 S. Ilhan, N. Duru, E. Adali, Improved Fuzzy Art Method for Initializing K-means, *International Journal of Computational Intelligence Systems*, **3**(3) (2010), 274–279.