A Prototype Application of Fuzzy Logic and Expert Systems in Education Assessment

James R. Nolan

Siena College
Departments of Quantitative Business Analysis and Computer Science
515 Loudon Road Loudonville, NY 12211 USA
jnolan@siena.edu

Abstract

This paper reports on the design and development of an expert fuzzy classification scoring system for grading student writing samples. The growing use of written response tests in the education sector provides fertile domain areas for new and innovative applications of soft computing and expert systems technology. The main function of the expert fuzzy classification scoring system is to support teachers in the evaluation of student writing samples by providing them with a uniform framework for generating ratings based on the consistent application of scoring rubrics. The system has been tested using actual student response data. A controlled experiment demonstrated that teachers using the expert fuzzy classification scoring system can make assessments in less time and with a level of accuracy comparable to the best teacher graders. The paper introduces fuzzy classification techniques that can encapsulate knowledge about imprecise qualities needed for constructing rule-based scoring models that provide consistent, uniform scoring results. This increased consistency in the application of the scoring rubrics allows for more valid individual and group assessment.

Introduction

A task central to all education programs is one of assessment. Assessment based on standardized tests is popular because it is thought that both individual and group comparisons will be possible if all participants take the same test and this test is graded the same way each time. The failure to perform *accurate* assessments in a timely manner may result in delays in providing student access to developmental activities needed to improve. In addition, students and school districts cannot be compared if the assessment process is not designed to be consistent and uniformly implemented.

The last several years have seen an increased emphasis on assessment of student writing ability. Many school systems and state education departments have designed writing assessment instruments or have contracted with outside testing centers to develop and grade what are deemed to be standard evaluative tests for student writing. In 1995 alone, the Educational Testing Service, the world's largest provider and grader of exams that require narrative responses, rated by hand some nine million pieces of student writing (Page & Petersen, 1996).

The task of grading student writing on standardized tests is very repetitive and labor intensive. Typically a teacher must learn a scoring standard or "rubric" that he or she will consistently apply to all student writing samples. Applying the scoring rubric consistently generally takes a considerable amount of time. In addition, the scoring rubrics for writing assessment usually employ the use of linguistic categories and approximate reasoning. This makes it much more difficult to ensure uniform application of the scoring rubrics. Expert decision support help in making the grade decision could lead to quicker evaluation of writing samples and more valid individual and group assessment because the application of the scoring rubrics would be much more uniform.

Several researchers have designed rubrics for performance-based assessment of a student's writing ability (Brewer, 1996; Marzano, et al, 1993). To date, no one has reported on efforts to create an automated scoring system employing rubrics for grading student writing samples. Since any type of grading is a classification task, the use of a classification model based on the particular scoring rubrics will help to standardize the grading (Ebert, 1996). The rubrics developed for grading student writing generally use linguistic categories, e.g., "thorough" understanding. The problem with linguistic categories is that they are imprecise or "fuzzy". The purpose of this paper is to introduce an expert fuzzy classifier model for grading student writing samples. It will be shown that the expert fuzzy classifier model can encapsulate rules using linguistic categories and results in faster, more consistent grade assignments. An overview of the application, including the problems with human scoring, follows. The next section reviews the theory of fuzzy classification. We apply this theory and describe the development of an expert fuzzy classification scoring system for grading student writing on standardized tests. Finally, the expert fuzzy classifier system is validated and conclusions drawn. Areas for further research are highlighted.

Overview of the Application

The New York City School District has developed a program called Curriculum Frameworks which stresses language arts at all grade levels. They subsequently developed the Performance Assessment in Language Arts (PAL) test to measure both reading and writing, integral parts of the language arts curriculum effort. In the New York City Grade 4 PAL, fourth grade students are asked to read a story and a poem and to respond to essay response questions about each.

All items in the NYC Grade 4 PAL test require students to generate individual narrative responses, rather than select a response from a list of choices. The response is evaluated by teachers who have been trained to exercise their professional judgment in applying scoring rules called rubrics. The rubrics describe the kind of student work that corresponds to the various score points. By using rubrics with staff development training, it is hoped that grading standards will be applied consistently across the city-wide school system.

The scoring guide to be used by all teachers in grading the Grade 4 PAL test is divided into two sections. The first section describes how to score the test for reading comprehension. The second part details the scoring rules for writing in response to literature. The general reading comprehension rubric is designed to aid the teacher in making a holistic judgment as to which category the student's response fits into: *high* (insightful, thoughtful), *medium* (basic, no frills), or *low* (confused, missing pieces). The specific reading comprehension rubric, as opposed to the general or "holistic" reading comprehension rubric, gives the teacher examples of specific ideas directly related to the story and tells which score category they fit into (0 through 6, inclusive). Finally, the reading comprehension anchor papers are specific examples of responses to the questions. These responses are provided for each score category.

The writing in response to literature section of the test has two parts. The first part is designed to test for writing effectiveness. This involves two dimensions: content and style. Each dimension has its own rubric. After employing each rubric, the teacher adds the scores together. The second part of the writing in response to literature section of the NYC Grade 4 PAL test concerns writing mechanics, i.e., spelling, grammar, etc.

Problems with Human Scoring

As is evident from this brief overview of the NYC Grade 4 PAL test and its scoring rules or rubrics, the teachers are trained to follow a scoring rubric that asks them to group the student's response into one of three general categories - high, medium, or low. After that they are to assign a numerical score by using the specific guidelines and rules established for each score level. Because the results of this test are to be used as evaluative measures for both individuals and groups, it is imperative that the student papers are scored the same way every time. As mentioned earlier, ensuring this to be true is problematic. Although the New York City School District has done a good job of developing scoring rules and standards, there is no way of ensuring that they are being applied the same way by different raters. The literature shows that two raters will agree with each other only 65% of the time (Page & Petersen, 1996). Even more important than this statistic is the fact that their is no way to prove that a particular teacher grading the NYC Grade 4 PAL test is applying the scoring rubrics in the same exact way each time. Fatigue and a myriad of other personal factors may affect consistency.

The second problem is time itself. Dutifully applying these well thought-out scoring rubrics to thousands of student papers every year takes a considerable amount of teacher time. There would be significant benefits if an expert decision support system could be developed that would serve to help the teacher in applying these scoring rubrics for classifying student writing. The scoring rubrics would be applied the same way every time and the scoring of many writing samples could be done in a more efficient way, leaving valuable time for the teacher to spend on developmental tasks rather than evaluative tasks.

Since the NYC Grade 4 PAL scoring rubrics are composed of rules with imprecise categories for making grade classifications, fuzzy logic and fuzzy classification models were used to represent these rubrics in a rule-based expert system environment. The theory of fuzzy classification is described next.

Fuzzy Classification

Fuzzy logic refers to a mode of reasoning in the presence of imprecise or ambiguous information (Zadeh, 1997). Fuzzy logic technology enables one to perform approximate reasoning and improves performance of classification systems in three ways: through efficient numerical representation of vague terms and concepts, by increasing their range of operation in ill-defined environments, and by decreasing their sensitivity to noisy data.

Membership Functions

Contrary to its nomenclature, fuzzy logic-based systems operate based on the precise and rigorous mathematics of fuzzy sets. A fundamental concept of fuzzy sets is that an element **x** may be a member of set **A** with varying degrees, i.e., each member of the set is characterized by its *degree of membership* within the set. The degree of membership of element **x** in set **A** is denoted by $\mu_A(\mathbf{x})$. A mapping of the domain interval to its degree of membership defines the membership function μ.

For example, different teachers have different definitions of what is a *medium* level of understanding in reading comprehension. In fact, most teachers would agree that no single value defines the category perfectly. If the student's understanding of a reading was being measured by adding the number of reading comprehension questions answered correctly, *medium* could apply to a range of total correct answers, with each number in the interval being *more medium* or *less medium* relative to some "typical" or "ideal" value. Assuming that 5 questions answered correctly out of ten is a typical *medium* level of understanding, a membership function for understanding might appear as a bell-shaped curve centered around 5 questions answered correctly. To accelerate computations and/or to reduce computer memory requirements, membership function shapes are usually simplified to triangular or trapezoidal forms.

Each linguistic variable is usually associated with a complete set of membership functions that are defined over the entire operating range of that variable. In fact, in fuzzy logic systems, neighboring membership functions overlap to indicate that a value may belong to different sets at the same time, with different degrees of membership. The number of membership functions assigned to the features to be used for classification, and the shape of these membership functions, comprise an essential part of the knowledge embodied in a fuzzy logic system. This information is usually supplied by the domain expert and, when combined with the rulebase(s), a complete knowledge-base for a particular application is formed.

Fuzzification

Fuzzification refers to the process of determining the degree of membership of a crisp input data value among a feature variable's membership function set. The "fuzzified" values are determined by intersecting the input value to the fuzzy set associated with each linguistic label. For instance, an input value of 6 correct answers out of 10 reading comprehension questions might result in a degree of membership in the set labeled "medium" of 0.7 and a degree of membership in the set labeled "high" of 0.3.

Fuzzy Rulebase

Fuzzy logic classification systems are typically implemented in the form of an expert decision support system. These expert systems make decisions and generate output values (classifications) based on the knowledge provided by the designer in the form of IF<*condition*> THEN<*action*> rules. The rulebase contains a collection of rules and forms an integral part of the total knowledge embedded in the system.

Rules are generally specified by the domain expert. The <*condition*> and <*action*> parts of each rule are denoted as the antecedent and consequent parts, respectively. The antecedent part can be a simple or complex logical combination of conditions, and more than one action may be specified in the consequent part. Examples of some fuzzy logic rules are:

IF understanding is *high* and character-
recognition is *strong*
THEN reading-comprehension is *high*
IF understanding is *low* and character-
recognition is *weak*
THEN reading-comprehension is *low*

During each pass of the fuzzy logic system operation, the rulebase is evaluated and outputs generated. There are many ways of computing the output of individual rules and combining the results (and resolving conflicts). They are termed fuzzy inference methods.

Fuzzy Rule Evaluation (Inference)

Fuzzy or approximate reasoning involves decision making based on ambiguous or ill-defined assumptions and incomplete data (Kasabov, 1996; Zimmermann, 1991). A fuzzy logic-based classification system generates output categories by inferring cause-and-effect relationships provided in its knowledge-base (the collection of rules and membership functions). The goal is to generate the most logical (best possible) conclusions even for rules with multiple antecedent conditions and rulebases with conflicting rules. The process of "inferring" conclusions based on certain assumptions and premises that are satisfied in whole or in part, is termed the inference process. Fuzzy inference entails the evaluation of rules, resolution of conflicts, and aggregation of multiple recommendations.

Many different inference strategies exist, and each method varies in the combination of processing techniques for dealing with conjunctions/disjunctions in the rule antecedent, inferring the antecedent grade of membership to output fuzzy sets, resolving conflicts from multiple rules on the same output membership function, and merging the contribution of different rules to the same output variable. We will discuss what is arguably the most popular fuzzy inference strategy, max-min inference. In the max-min inference method, the min operation is used for the AND conjunction (set intersection) and the max operation is used for the OR disjunction (set union) in order to evaluate the grade of membership of the antecedent clause in each rule. For example, assume a student answers 6 out of 10 correct on the reading comprehension questions. Suppose fuzzification for the variable *understanding* produces a 0.7 degree of membership in the set "medium" and a 0.3 degree of membership in the set "high". Additionally, assume a student scores 3 out of 5 on the character recognition questions and fuzzification for the variable *character-recognition* produces a 0.6 degree of membership in the set "strong" and a 0.4 degree of membership in the set "weak", then:

RULE 1: IF understanding is *medium* and
    character-recognition is *weak*
THEN reading-comprehension is *medium*
EVALUATION: min (0.7, 0.4) = 0.4 reading-
    comprehension is *medium*
RULE 2: IF understanding is *high* and
    character-recognition is *strong*
THEN reading-comprehension is *high*
EVALUATION: min (0.6, 0.3) = 0.3 reading-
    comprehension is *high*

The value 0.4 is used to "clip" the *medium* reading-comprehension output membership function shape. Similarly, the value 0.3 is used to "clip" the reading-comprehension output membership function shape for *high*. If multiple rules have the same consequent label, the max operation is used to resolve conflicts. The clipped membership functions resulting from the application of many rules are then merged to produce one final fuzzy set. The max operation is used to merge overlapping regions.

Defuzzification

When the inference process is complete, the resulting data for each output of the fuzzy classification system is a collection of fuzzy sets or a single, aggregate fuzzy set. The process of computing a single number that best represents the outcome of the fuzzy set evaluations is called defuzzification. There are many methods that can be used for defuzzification. The centroid, also referred to as the "center-of-gravity method, produces "crisp" output data by computing the horizontal-axis component of the geometric centroid of the fuzzy set. Intuitively, the centroid method can be viewed as a "compromise" among the output actions recommended by different rules. For each output using this defuzzification method, the resultant fuzzy sets from all contributed rules are merged into a final aggregate shape. The centroid of the aggregate shape is then computed.

Applying Fuzzy Classification to Student Writing

This section describes the development of an expert fuzzy classification scoring system for scoring student writing samples. The scoring rubrics for the NYC Grade 4 PAL test have many of the characteristics of fuzzy logic that were described earlier. Classifying student writing samples from this test involves reasoning in the presence of imprecise or ambiguous information. For example, the NYC Grade 4 PAL test scoring rubric

IF student demonstrates *high* level of
    understanding of the whole work
AND exhibits a *strong* level of recognition of
    important characters
AND recognizes *crucial* elements of the plot
AND generates *new* ideas
THEN rate reading comprehension *high*

involves the use of vague linguistic terms and concepts. Employing the fuzzy logic notion of degree of membership for each of the four underlined variables included in the antecedent (the IF portion) of the scoring rule cited above would allow us to "fuzzify" the data now used by the teachers to measure these terms. Identifying the possible categories for the variable in the consequent part of the scoring rubric will allow us to "defuzzify" the variable and obtain the final numeric score for reading comprehension. Doing this for all scoring rules will result in a knowledge-base similar to the one used by teachers who are trained to apply the scoring rubrics designed for the NYC Grade 4 PAL test.

Prototype Development

The development of the expert fuzzy classification system for scoring student writing samples proceeded in the following manner:

1. A group of expert teacher graders was selected and asked to develop ranges of scores corresponding to labels for each of the linguistic feature categories used in the scoring rubrics.
2. This "membership" data was used to develop membership function sets for each feature and classification variable.
3. A fuzzy rulebase was developed using the NYC Grade 4 PAL test scoring rubrics. This, along with the membership functions, comprised the knowledge-base of the expert fuzzy classification scoring system.
4. The expert fuzzy classification scoring system was validated by classifying test data sets.

The work resulted in the development of 21 membership function sets representing the feature and classification variables. The rulebase consists of over 200 rules. The teacher decides on the ratings to be given for each of the input feature variables, e.g., understanding, recognition of important characters, etc.. These ratings are automatically "fuzzified" and the appropriate rules from the rulebase are fired. The results are "defuzzified", resulting in numeric scores. The output from the classification component can be visualized and further explained by providing the underlying rules used to make the classification.

A commercially available software package called O'Inca Design Framework was used for developing the membership functions and rules. This fuzzy logic and expert system shell software package has additional facilities for simulation, on-line modification of rules and membership functions, and displaying output classifications and inference paths.

Testing and Validation

Two schools in New York City School District Six were selected as test sites. All grade 4 PAL tests completed by fourth grade students in these two schools were scored by teachers using the expert fuzzy classification scoring system. Over a one month period, 255 student writing samples were evaluated. At the end of the one month testing period, expert teacher graders from outside these schools reviewed the exams scored with the expert fuzzy classification scoring system and unanimously agreed the results of the scoring demonstrated consistent use of the rubrics designed for scoring the test. The teachers who used the system remarked about the speed with which they were able to evaluate the writing samples. They attributed this decrease in time needed to grade exams to relief of the burden of having to perform the mechanics of scoring them. They felt it enabled them to concentrate on evaluating the factors that are important in the holistic scoring method without having to worry about the actual manipulation of score categories.

A controlled experiment was set up to determine just how effective teachers evaluating student writing samples with the expert fuzzy classification scoring system were compared to domain experts (the expert teacher graders). Two hundred student writing samples were selected for the experiment. The three expert teacher graders reviewed each of the 200 writing samples and made an evaluation using the holistic scoring rubrics. The same 200 writing samples were independently reviewed and assessed by three different teachers using the expert fuzzy classification scoring system. The results indicated that the teachers using the expert fuzzy classification scoring system agreed with the three domain experts in 178 of the 200 cases for an agreement rate of 89%. Since it is not unusual for two teachers to disagree on the score to be assigned to the same sample of student writing, most standardized writing exams allow a difference of one point between the two graders before a compromise must be reached. If we use this criteria, 194 of the 200 cases would be considered in agreement (97%). There also was a significant difference in the time each group needed to do the scoring (see Tables 1 and 2).

Table 1
Comparison of Grade Classifications
Expert Teacher/Grader Assigned Scores

		0	1	2	3	4	5	6
	0	8						
Scores Assigned	1	1	23	2				
By Teachers	2	1	1	39	2	1		
Using Expert	3		1	1	38	1	1	
Fuzzy Classifi-	4			1	3	34	4	
cation Scoring	5				1	2	18	1
System	6						6	10

Correct Classification Rate: 89%

Table 2
Average Evaluation Time

Grader:	Time (min.)
Expert Teacher/Graders	15
Grader 1 (with system)	11
Grader 2 (with system)	9
Grader 3 (with system)	9

Costs

The costs associated with the development and testing of the prototype expert fuzzy classification scoring system have totaled approximately $65,000 so far. This includes the cost of hardware, software, and teacher/staff time. The estimated cost for further development and implementation in New York City School District Six is $40,000. The District is committed to full implementation.

Conclusions and Further Research

The use of fuzzy classification in an expert system environment has proven to be of value in the domain of scoring student writing samples. The problems that arise in the management of uncertainty and vagueness in the scoring of student writing samples have been discussed. Fuzzy logic provided a natural conceptual framework for representation of the imprecise knowledge and inference processes associated with the scoring process. The benefits of the expert fuzzy classification system are:

1. A significant reduction in the time it takes to score the standardized New York City Grade 4 PAL exam.
2. Increased consistency in the application of the scoring rubrics.
3. The system enables less experienced teachers to become more familiar with the scoring rubrics.

The impact of the expert fuzzy classification scoring system on the time it takes teachers to score a student writing sample is important. As discussed previously, one of the problems with using standardized writing sample evaluations is that they are time consuming. By reducing the time for scoring a student writing sample by approximately one third, the writing sample evaluation process becomes more efficient.

This increase in efficiency would not be valuable if the accuracy of the evaluation suffered. The test results show that the accuracy and consistency of the evaluation performed by teachers using the expert fuzzy classification scoring system is equal to that of the expert teacher graders. Finally, the newer, less experienced teachers who took part in the testing have remarked about the usefulness of the system for both learning the scoring rubrics and providing explanations of the grading. They found the explanation of the grading useful for designing developmental work for the students.

Future Research

Rule-based fuzzy classification expert systems offer the potential for new and more powerful applications of AI in all areas of assessment. The direction for future research is to generalize this approach to other areas of assessment, e.g., portfolio assessment, evaluation of proposals by funding agencies, etc. A limitation of the current system is that teachers must still read and provide input on the student's written response to the standardized exam questions. A long range research goal is to develop a "front-end" to the expert fuzzy classification scoring system that will provide the inputs now given by the teacher.

REFERENCES

Brewer, R. (1996). *Exemplar's A Teacher's Solution.* Underhill, VT: Exemplar.

Ebert, C. (1996). Fuzzy classification for software criticality analysis. *Expert Systems with Applications,*11(3),323-342.

Kasabov, N. K. (1996). Foundations of neural networks, fuzzy systems, and knowledge engineering. Cambridge, MA: MIT Press.

Marzano, R., Pickering, D. & McTighe, J. (1993). *Assessing Student Outcomes: Performance Assessment Using the Dimensions of Learning Model.* Alexandria, VA: ASCD.

Page, E. B., & Petersen, N. S. (1996). The computer moves into essay grading: updating the ancient test. *Phi Delta Kappan*, March, 561-565.

Zadeh, L. (1997). In Jang, J. S., Sun, C. T., & Mizutani, E. (Eds.). *Neuro-fuzzy and soft computing: A conceptual approach to learning and machine intelligence*, Upper Saddle River, NJ: Prentice Hall.

Zimmermann, H. G. (1991). *Fuzzy set theory and its applications*, 2nd ed., Boston, MA: Kluwer.