JHM
JOURNAL OF HONAI MATH

p-ISSN: 2615-2185
e-ISSN: 2615-2193

Journal of Honai Math, Vol. 2, No. 2, pp. 77-84, Oktober 2019
JOURNAL OF HONAI MATH
http://journalfkipunipa.org/index.php/jhm
http://doi.org/10.30862/jhm.v2i2.81

EDUCATIONAL FUZZY DATA-SETS AND DATA MINING IN A LINEAR FUZZY REAL ENVIRONMENT

Frank Rogers
Department of Mathematics, University of West Alabama, Livingston, Alabama 35470, USA
Email: frogers@uwa.edu

Abstrak

Penambangan data pendidikan merupakan proses mengubah data mentah dari sistem pendidikan ke informasi yang bermanfaat yang dapat digunakan oleh pengembang perangkat lunak pendidikan, siswa, guru, orang tua, dan peneliti pendidikan lainnya. Kumpulan data pendidikan fuzzy adalah kumpulan data yang terdiri dari nilai yang tidak pasti. Tujuan dari penelitian ini adalah untuk mengembangkan dan menguji model klasifikasi di bawah ketidakpastian yang unik untuk siswa modern. Ini dilakukan dengan mengembangkan model data tidak pasti yang berasal dari lingkungan pendidikan dengan data Linear Fuzzy Real. Mesin pembelajaran kemudian digunakan untuk memahami siswa dan lingkungan belajar yang optimal. Kemampuan untuk memprediksi kinerja siswa adalah penting dalam lingkungan online atau berbasis web. Hal ini berlaku juga di kelas yang ideal dan secara khusus penting untuk daerah pedesaan di mana prestasi akademik yang kondisinya jauh dari ideal.

Kata kunci: Lingkungan Belajar Siswa, Kumpulan Data Pendidikan Fuzzy, Bilangan Ril Fuzzy Linier, Mesin Pembelajaran.

Abstract

Educational data mining is the process of converting raw data from educational systems to useful information that can be used by educational software developers, students, teachers, parents, and other educational researchers. Fuzzy educational datasets are datasets consisting of uncertain values. The purpose of this study is to develop and test a classification model under uncertainty unique to the modern student. This is done by developing a model of the uncertain data that come from an educational setting with Linear Fuzzy Real data. Machine learning was then used to understand students and their optimal learning environment. The ability to predict student performance is important in a web or online environment. This is true in the brick and mortar classroom as well and is especially important in rural areas where academic achievement is lower than ideal.

Keywords: Student Learning Environment, Educational Fuzzy Data-sets, Linear Fuzzy Real Numbers, Machine Learning

INTRODUCTION

The GDP of a country is an indicator of economic health (Well, 2007; Stuckler, et al. 2009). An indirect reflection of this health is the education level of a countries' citizenry. By presenting quality education to students on every level, educators along with

other educational stakeholders are contributing to the health and wealth of a nation (Minikwa, 2017). One way to accomplish an increase in the quality of education is by predicting student's performance and taking early actions to improve the student's ability. The data needed to predict student's performance lies within both crisp and fuzzy datasets (Cheng, 2017).

Using fuzzy data mining techniques, educators can apply fuzzy machine learning and analytical techniques to classify students and to better predict the performance of students. This will allow a viable method of intervention unique to the student. Fuzzy decision making was initially introduced by Bellman and Zadeh (1970). This concept was then adopted to mathematical programming, uncertain programming, fuzzy partially ordered sets (Neggers & Kim, 2001), linear programming (Dubois & Prade, 1982) and many other concepts.

Linear Fuzzy Real (LFR) numbers are a system of numbers that have properties unique to the set of real numbers and the set of fuzzy numbers (Rogers, 2016). LFR numbers are used in the study of fuzzy random variables (Monk, 2001). Rogers (in press) stated that it was demonstrated that gradient-based machine learning algorithms in a linear fuzzy real environment performed specific task effectively without using explicit instructions.

Linear Fuzzy Real Numbers

This section describes the linear fuzzy real numbers. Fuzzy real numbers are hybrid fuzzy numbers that exhibit the properties of both real numbers and traditional fuzzy numbers (Monk, 2001; Prevo, 2002; Rogers, 2016). Considering the set of all real numbers R, one way to associate a fuzzy number with a fuzzy subset of real numbers is as a function μ: $R \rightarrow [0,1]$, where the value $\mu(x)$ is to represent a degree of belonging to the subset of R.

Definition 1. *(Linear fuzzy real number) Let* μ: $R \rightarrow [0,1]$ *be a function such that:*

1. $\mu(x) = 1$ *if* $x = b$;
2. $\mu(x) = 0$ *if* $x \leq a$ *or* $x \geq c$;
3. $\mu(x) = (x - a)/(b - a)$ *if* $a < x < b$;
4. $\mu(x) = (c - x)/(c - b)$ *if* $b < x < c$.

Then $\mu(a,b,c)$ is called a linear fuzzy real number with associated triple of real numbers (a,b,c) where $a \leq b \leq c$ shown in Figure 1.

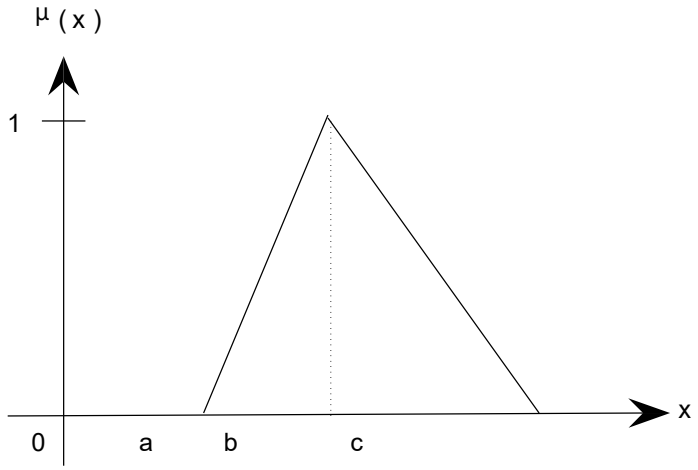

Figure. 1. Linear fuzzy real number $\mu(a,b,c)$

Given that *LFR* is the set of all linear fuzzy real numbers, any real number b can be written as a linear fuzzy real number, $r(b)$, where $r(b) = \mu(b,b,b)$ and so $R \subseteq LFR$. In this paper, $b = \epsilon(b) = r(b)$ represents the real number b itself. Arithmetic operations, algebraic operations, functions, linear equations and nonlinear equations are also defined in LFR (Rogers, 2016).

Classification Methods

Classification is one of two types of supervised machine learning problems. Supervised machine learning involves fitting a model that relates independent variables to the dependent variables. The dependent variables are typically denoted as the predictors while the independent variables are denoted as the responders. In supervised learning the aim is to fit the model to allow the user to accurately predict the response for future observations or better understanding the relationship between the dependent and independent variables (Muller & Guido, 2016).

In classification, the dependent variables are class labels and thus the goal is to predict the class label, from a predefined list of possibilities. A binary classification distinguishes between exactly two classes. The other supervised machine learning problem is the regression problem. The Classification problem occurs often, perhaps more

often than the regression problem. And while the output variable in regression is numerical, the output variable for classification is categorical (Muller & Guido, 2016). Three of the most widely-used classifiers are logistic regression, linear discriminant analysis and K-nearest neighbors. The Linear Fuzzy Real logistic regression classifier was used for the purpose of this research. Setting the binary classification to dummy variables 0 and 1, allows the use of the logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

The maximum likelihood along with fuzzy inputs was used to fit the Linear Fuzzy Real logistic classifier. Thus, it is possible to minimize the cost function using the modified gradient descent (Rogers, in press). The fuzzy input addresses real-world applications that resemble human decision making (Bellman & Zadeh, 1970). Therefore, the purpose of this study is to determine the viability of using Machine Learning techniques on a Linear Fuzzy Real Dataset to understand students and their optimal learning environment.

METHOD

A Likert survey was given to 172 students belonging to four elementary schools in the Blackbelt of Alabama and Mississippi, United States of America. The Blackbelt is a region in the southern area of the United States of America that encompasses parts of Mississippi, Alabama and several other southern states. The rural communities of the Blackbelt have historically faced serious poverty, poor health care and inadequate education programs (Ballestros, 2017). For each student, their instructor recorded an evaluation of the student. Student's identities were withheld. The teacher's evaluation served as the dependent variable. The student's survey answers served as the fuzzy independent variables.

The questions focused on activities that the elementary student or perhaps the parent could control. Questions about the student's sleep habits, exercise habits, parent involvement, formed opinions about reading and arithmetic, and student's academic confidence serve as the dependent variable. The survey contained thirteen questions. However, only ten represented significant independent variables. Again, the independent variable is the teacher evaluation of each individual student. Figure 2 illustrates the first four questions of the survey.

Instruction for Students: Please choose the answer closest to what you believe is true
1. When my teacher is explaining the lesson, I...
A. Never Take notes
B. Take Notes Sometimes
C. Always Take Notes
2. When preparing for a test
A. I don't study at all.
B. I study the night before the test.
C. I study more than three days before the test.
3. I think math
A. Is boring
B. Interesting
C. Kind of fun
4. I sleep for
A. Around 6 hours or less
B. 6 to 8 hours
C. More than 8 hours

Figure 2. Sample survey questions

The Figure 3 shows the sample data of the first four questions, set for predicting third and fourth grade results for the first fifteen students from the Black Belt of Alabama and Rural Central Mississippi.

Student	Question1	Question2	Question3	Question4	Question4
1	2	1	3	mu(3,3,3.8)	3
2	2	2	2	mu(3,3,3.8)	3
3	2	2	2	mu(3,3,3.8)	3
4	2	2	2	mu(3,3,3.8)	3
5	2	2	3	mu(1.2,2,2.8)	2
6	2	2	2	mu(3,3,3.8)	3
7	3	1	1	mu(0.2,0.2,1)	1
8	2	2	2	mu(1.2,2,2.8)	2
9	2	1	3	mu(0.2,0.2,1)	1
10	2	2	3	mu(3,3,3.8)	3
11	2	2	2	mu(0.2,0.2,1)	1
12	2	2	3	mu(3,3,3.8)	3
13	2	3	2	mu(3,3,3.8)	3
14	1	1	2	mu(1.2,2,2.8)	2
15	2	3	3	mu(3,3,3.8)	3

Figure 3. Sample Data

The values, 1, 2, and 3 are used to represent the quantitative values, A, B, and C respectively. Question 4's inputs are one of two Linear Fuzzy Real inputs in the data. The

Linear Fuzzy Real numbers are used to represent human ambiguity. The elementary teachers were essential to the research. The choice of which independent variables are fuzzy or crisp as well as important to the academic health of the students was a consequence of informed conversations with the elementary instructors.

An initial test of the fuzzy machine learning algorithm was also done, using data obtained from an online survey. The survey was given online to first-year college students. The independent variables were a set of questions on the survey. The dependent variable is a set of data on the survey designed to measure the student's academic confidence. The values for the academic confidence were averaged out to obtain a set of continuous independent values.

RESULTS AND DISCUSSION

Data is critical when using Machine learning techniques. Possible areas of concern are dependency of the independent variables, accuracy of the model, and number of entries (Gareth, et al. 2015). The focus of this paper is using machine learning techniques upon elementary fuzzy educational data sets; however, the concept applies to all levels of education. For this reason, experimenting with college level students proved informative.

First Year College Student Data

The college data consisted of the online surveys that targeted freshmen-level college students. It was imported into Excel to be cleaned and observed. Once the data had been cleaned, it was then opened in Python as a data frame. The data was split into a testing and training set. The training set consisted of 345 entries and 10 columns. The average for the student academic confidence was the cutoff for student success. The classifier, Logistic Regression, had an intercept scaling of 1, tolerance of 0.001 and a max iteration of 100,000. The purpose of the college experiment was to test the effectiveness of the survey as well as the effectiveness and accuracy of the Linear Fuzzy Real Logistic Regression algorithm. For this experiment, a Linear Fuzzy Real accuracy rate of around 93 percent, was obtained.

Elementary School Data

Our elementary dataset was first imported into Excel to be cleaned and observed directly. Once the data had been cleaned, it is then opened in Python as a data frame. The data was split into a testing and training set. The training set consisted of 129 entries and

ten columns. The dependent variable consisted of students who were at risk, below a c-average. And those who were doing well, c-average and above.

The classifier, Linear Fuzzy Real Logistic Regression, had an intercept scaling of 1, tolerance of 0.001 and a max iteration of 100,000. For this experiment, a Linear Fuzzy Real accuracy rate of around 90.7 percent, was obtained. Written as a discrete LFR, the fuzzy accuracy rate is $\mu(88.4, 90.7, 90.7)$ percent. This implies that the accuracy is bounded between 88.4 and 90.7, with 90.7 the likeliest accuracy rate. While only two of the independent variables was written as LFR numbers, the logistic regression algorithm produced LFR coefficients for each variable.

CONCLUSSION

Successful classification at 90% or above, suggest that a student's success could be predicted using the model. The model itself is fitted for at-risk students, the educators involved in this study, define students as at-risk if they are in danger of obtaining a grade below average. The definition of average is user-defined; however, the educators chose a grade of 70% on a 100-point scale.

For the purpose of our dummy variables, a classification of 0 was given to students with a designated score of less than 70% and a classification of 1 was given to students with a score of 70% or more. It is noted that the classification variables can be switched, where a classification of 1 is given to students with a score below 70% and a classification of 0 is given to students who are 70% or above, and the results will essentially be the same. However, it can be adjusted to classify students in any categories such as excellent or outstanding. This would be done to classify gifted students. This is however dependent upon the needs of each classroom. The data as well as the coefficients found may be unique to the schools involved. A collection of data, perhaps distinguished by region, state or even district, can be gathered and mined to classify their students and their unique circumstances. With data unique to the schools and the prediction of student's academic performance, an optimal environment can be created for a group of students or for the individual student.

REFERENCES

Ballestros, C. (2017). "Alabama has The Worst Poverty in the Developed World, U.N. Official Says," *Newsweek*, accessed at https://www.newsweek.com/alabama-un-

poverty-environmental-racism-743601.

Bellman, R.E., & Zadeh, L.A. (1970). Decision making in a fuzzy environment. *Management Science*, *17*(4), B141–B164.

Cheng, J. (2017). Data-mining research in education. *Report.* Hongshan: International School of Software, Wuhan University.

Dubois, D., & Prade, H. (1982). System of linear fuzzy constraints. *Fuzzy Sets and Systems, 13*, 1–10.

Gareth, J., Witten, D., Hastie, T., & Tibshirani, R. (2015). *An Introduction to Statistical Learning*. New York: Springer.

Minikwa, N. (2017). *Principle of Macroeconomics*. Boston: Cengage Learning.

Monk, B. (2001). A proposed theory of fuzzy random variables. *Dissertation*. Tuscaloosa: University of Alabama.

Muller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists*. California: O'Reilly Media.

Neggers, J. & Kim, H. (2001). Fuzzy possets on sets. *Fuzzy Sets and Systems*, *117*(3), 391-402.

Prevo, R. (2002). Entropies of families of fuzzy random variables: an introduction to an in-depth exploration of several classes of important examples. *Dissertation*. Tuscaloosa: University of Alabama.

Rogers, F. (2016). Linear fuzzy integers and Bezout's identity. *Asian Journal of Fuzzy and Applied Mathematics, 4*(2), 5-10.

Rogers, F. (in press). Fuzzy gradient descent for the linear fuzzy real number system. *AIMS Mathematics*.

Stuckler, D., Basu, S., Suhrcke, M., Coutts, A., & McKee, M. (2009). The public health effect of economic crises and alternative policy responses in Europe: An empirical analysis. *The Lancet*, *374*(9686), 315-323.

Well, D. N. (2007). Accounting for the effect of health on economic growth. *The Quarterly Journal of Economics*, *122*(3), 1265-1306.