Predicting students' grades using fuzzy non-parametric regression method and ReliefF-based algorithm

Javad Ghasemian[1], Mahmoud Moallem[1] and Yasin Alipour[2]

[1] School of Mathematics and Computer Sciences, Damghan University

Damghan, Iran

ghasemian@du.ac.ir, moallem@du.ac.ir

[2] Information Technology Department, Damghan University

Damghan, Iran

alipour@du.ac.ir

Abstract

In this paper we introduce two new approaches to predict the grades that university students will acquire in the final exam of a course and improve the obtained result on some features extracted from logged data in an educational web-based system. First we start with a new approach based on Fuzzy non-parametric regression; next, we introduce a simple algorithm using ReliefF estimated weights. The first prediction technique is yielded by integrating ridge regression learning algorithm in the Lagrangian dual space. In this approach, the distance measure for fuzzy numbers that suggested by Diamond is used and the local linear smoothing technique with the cross validation procedure for selecting the optimal value of the smoothing parameter is fuzzified to fit the presented model. Second approach is based on ReliefF attribute estimation as a weighting vector to find the best adjusted results. Finally, to check the efficiency of the new proposed approaches, the most popular techniques of traditional data mining methods are compared with the presented methods.

Keywords: *Educational Data Mining, Predicting Marks, Fuzzy Non-parametric Regression, KDD, ReliefF, WEKA, Matlab*

1. Introduction

Globally the application of data mining [25] in education is great. Educational data mining (also known as EDM) is a type of knowledge discovery science and focused on the development of techniques for making discoveries within the unique types of data that gathered from educational environments, and using those methods to understand efficaciously the students and help them to learn better and potentially ameliorate some aspect of educations. These data can be extracted from a number of sophisticated web-based learning and course management tools called Virtual Learning Systems (VLSs) such as Moodle, eFront, ATutor and many others, include among other features such as course content delivery features, assignment submission, online conferences, quiz modules, grade reporting system and log books [1].

Educational Data Mining can be used in many aspects of education, from students, to instructors, as well as staff to improve teaching/learning process and make better decisions about educational activities. Hence, the prediction of student performance with high accuracy is useful in many contexts in all educational institutions for identifying slow learners and distinguishing students with low academic achievement or weak students who are likely to have low academic achievements.

The increase of transactional educational systems as well as rich databases of student information has created large repositories of valued data reflecting how students learn. On the other hand, the use of internet in education has created a new context known as e-learning or web-based education in which large amounts of information about teaching/learning interactions are endlessly generated and ubiquitously available. All this information provides a gold mine of educational data [2] and also makes some challenges for researchers for a long time.

In this paper we have presented two new approaches. First we introduced a novel approach based on fuzzy non-parametric regression by integrating ridge-type regularization in the Lagrangian dual space and using Gaussian kernel as well as smoothing parameter, all together to reach the accurate prediction. And second approach introduced an algorithm to integrate a ReliefF weighting criteria as a weighting vector to find the adjusted result. In both presented proposals, with discovering of dataset composed of crisp inputs, we can infer a single aspect of data (student marks) from some combination of other aspects of data (such as online quizzes).

In the next section, we summarize the related work. The proposed statistical approach includes concepts and applied methods introduced in section 3. Section 4 introduces of integrating ReliefF weighting criteria as a weighting vector to use in prediction. Section 5 introduces the case study and the data in this study carried out. Section 6 reports on and compares experimental results for all algorithms tested. Finally, Section 7 concludes the paper.

2. Related work

In last decades, many studies around the determining of students' performance have been performed. Many statisticians have tried to predict and examine the outcomes [3] and many educational psychologists have tried to understand and explain the issue [4].

Also with the widespread accessibility of the World Wide Web services and evidence of e-learning solutions many technological approaches have been emerged. Naeimeh Delavari et al. [5], proposed a model with different types of education-related guidelines and the data mining techniques appropriate for dealing with large amounts of generated data in higher learning institutions. Luan et al. [6], express an instance of a specific case study of clustering students with similar characteristics. Zafra and Ventura [7] proposed an innovative technique based on multi-objective grammar guided genetic programming to detect the most relevant activities that a student needs to pass a course in virtual learning environment. Chanchary et al. [8], analyze student logs belong to a learning management system with data mining and statistical tools to discover relationships between student's access behavior and overall performance. Fausett and Elwasif [9], predict student's grades (classified in five classes: A, B, C, D and E or F) from test scores using neural networks. Martnínez [10], predicts student academic success (classes that are successful or not) using discriminant function analysis. Minaei Bidgoli and Punch [11], classify students by using genetic algorithms to predict their final grades. Kotsiantis and Pintelas [12], predict a student's marks (pass and fail classes) using regression techniques for the students at Hellenic Open University. Romero et al. [13] show how web usage mining can apply in e-learning systems in order to predict the final marks and Amelia Zafra, et al. [14] presented a multiple instance learning for classifying students for data in Cordoba University.

As a valued reference, Romero and Ventura [15] provided a survey which contains a categorized review of the main research studies using educational data mining techniques in the virtual learning environments.

3. First Approach: Fuzzy Non-parametric Regression methods

Generally there are two ways to develop a fuzzy regression model: (1) models where the relationship of the variables are fuzzy; and (2) models where the variables themselves are fuzzy. In order to formulation the problem, we focus on models, in which the data and relationship between variables are fuzzy.

Although many practical situations has implemented using parametric forms of fuzzy regression, large datasets with a complicated underlying variation trend, that used fuzzy parametric regression, may have produce unrealistic outcomes. Hence, some other approaches have been developed to deal with the fuzzy non-parametric regression problems; such as Ishibuchi and Tanaka [16], that integrated several fuzzy non-parametric regression techniques with traditional back-propagation networks and Cheng and Lee [17], have applied the radial basis function in fuzzified neural networks. Also with respect of significant development of statistical non-parametric smoothing methods, integrating smoothing techniques into non-parametric regression problems lead to achieve the better prediction.

In this study, we concentrate on ridge regression method, which integrated with non-parametric local linear smoothing (LLS), that is a special case of the local polynomial smoothing technique, which is fuzzified to handle fuzzy non-parametric regression with triangular fuzzy numbers based on the distance measure proposed by

Diamond [18]. Figure 1 shows a representation of triangular fuzzy number. A distance based on cross validation procedure for selecting the optimal value of the smoothing parameter is also suggested.

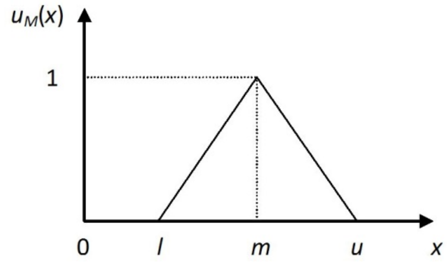

Fig. 1 A triangular fuzzy number M

In the rest of the study first explain the basic concepts of triangular fuzzy numbers and the local linear smoothing method. Next fuzzy ridge non-parametric regression model will be presented. Then appropriate kernel function and the smoothing parameter have been determined.

3.1 Basic concepts

As described later, this section is focused on fuzzy non-parametric regression model with multiple crisp input and triangular fuzzy output. In this section, based on local linear smoothing approach, a fitting procedure is proposed for this model.

Assume $a = (m_a - \alpha_a , m_a, m_a + \beta_a)$ be a triangular fuzzy number with its left, center, and right spread being. The membership function of a is:

$$\mu_a = \begin{cases} \dfrac{t-(m_a - \alpha_a)}{\alpha_a} & \text{if } m_a - \alpha_a \le t < m_x \\ \dfrac{m_a + \beta_a - t}{\beta_a} & \text{if } m_a \le t < m_a + \beta_a \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

In this paper we point to the space of all fuzzy triangular numbers by T(R), i.e.:

$$T(\mathrm{R}) = \{a: a = (m_\alpha - \alpha_\alpha, m_\alpha, m_\alpha + \beta_\alpha)\} \qquad (2)$$

Now, look at the following multi variable fuzzy non-parametric regression model:

$$Y = F(x)\{+\}\varepsilon = \big(m(x) - \alpha(x), m(x), m(x) + \beta(x)\big)\{+\}\varepsilon \qquad (3)$$

In this model, $x = (x_1, x_2, \dots, x_p)$ is a p-dimensional crisp independent variable (input) where, its domain is assumed to be $D \subseteq R^p$; also $Y \subseteq T(R)$ is a triangular fuzzy dependent variable (output). F(x), a mapping from D to T(R), is an unknown fuzzy regression function with its center, lower and upper limits being respectively $m(x)$, $l(x) = m(x) - \alpha(x)$ and $r(x) = m(x) + \beta(x)$. Moreover ε is an error term. Instead of being solely regarded as a random error with mean zero, ε may also be considered as a fuzzy error or a hybrid error containing both fuzzy and random components. {+} is an operator whose definition depends on the fuzzy ranking method used.

3.2 Local Linear Smoothing Method

Let $(x_i, Y_i)| \ x_i \in R^p, i = 1,2,3, \dots, n$ be a sample of the observed crisp inputs and triangular fuzzy outputs of model (1) with the underlying fuzzy regression function: $F(x) = (l(x), m(x), r(x))$. The main object in fuzzy non-parametric regression is to estimate F(x) at any $x \in D \subseteq R^p$ based on $(x_i, Y_i), i = 1,2, \dots, n$. As pointed out by Kim and Bishu [19], the membership function of an estimated fuzzy output should be as close as possible to the corresponding observed fuzzy number. From this point of view, we shall estimate $m(x), l(x)$ and $r(x)$ for each $x \in D$ in the sense of best fit with respect to some distances that can measure the closeness between the membership functions of the estimated fuzzy output and the corresponding observed one. Suppose that $m(x)$, $l(x)$ and $r(x)$ have continues partial derivatives with respect to each component x_i in the domain D of x. Then, for a given $x_0 = (x_{01}, x_{02}, \dots, x_{0p}) \in D$ and with Taylor's expansion, $m(x)$, $l(x)$ and $r(x)$ can be locally approximated in a neighborhood of x_0, respectively by the following linear functions:

$$\begin{cases} l(x) \approx \tilde{l}(x) = l(x_0) + l^{(x_1)}(x_0)(x_1 - x_{01}) + \\ \qquad \dots + l^{(x_p)}(x_0)\big(x_p - x_{0p}\big), \\ m(x) \approx \tilde{m}(x) = m(x_0) + m^{(x_1)}(x_0)(x_1 - x_{01}) + \\ \qquad \dots + m^{(x_p)}(x_0)\big(x_p - x_{0p}\big), \\ r(x) \approx \tilde{r}(x) = r(x_0) + r^{(x_1)}(x_0)(x_1 - x_{01}) + \\ \qquad \dots + r^{(x_p)}(x_0)\big(x_p - x_{0p}\big), \end{cases} \qquad (4)$$

Where $l^{(x_j)}(x_0)$, $m^{(x_j)}(x_0)$, and $r^{(x_j)}(x_0)$, $j = 1,2,3, \dots, p$ are respectively the derivatives of $m(x)$, $l(x)$ and $r(x)$ with respect to x_j at x_0.

Let $a = (l_a, m_a, r_a)$ and $b = (l_b, m_b, r_b)$, $m_a, r_a, m_b, r_b \ge 0$ be any two triangular numbers in T(R). Diamond defined a distance between a and b as follows:

$$d(a, b)^2 = (l_a - l_b)^2 + (m_a - m_b)^2 + (r_a - r_b)^2 \qquad (5)$$

The distance (5) measures the closeness between the membership functions of two triangular fuzzy numbers. We henceforth based on this distance we can extend the local linear smoothing technique to fit the fuzzy non-parametric model (1). With the observed data

$$(x_i, Y_i) = \left(x_{i1}, \dots, x_{ip}, (l_{yi}, m_{yi}, r_{yi})_{T(R)}\right),$$
$$i = 1,2,3,\dots,n \qquad (6)$$

and based on Diamond's distance (5), the following locally weighted least-squares is formulated. That is

Minimize

$$\sum_{i=1}^{n} d^2\left((l_{yi}, m_{yi}, r_{yi})_{T(R)}, \left(\tilde{l}(x_i), \tilde{m}(x_i), \tilde{r}(x_i)\right)_{T(R)}\right) K_h(\|x_i - x_0\|) = \sum_{i=1}^{n}\left(l_{yi} - l(x_0) - \sum_{j=1}^{p} l^{(x_j)}(x_0)(x_{ij} - x_{0j})\right)^2 K_h(\|x_i - x_0\|)$$
$$+ \sum_{i=1}^{n}\left(m_{yi} - m(x_0) - \sum_{j=1}^{p} m^{(x_j)}(x_0)(x_{ij} - x_{0j})\right)^2 K_h(\|x_i - x_0\|)$$
$$+ \sum_{i=1}^{n}\left(r_{yi} - r(x_0) - \sum_{j=1}^{p} r^{(x_j)}(x_0)(x_{ij} - x_{0j})\right)^2 K_h(\|x_i - x_0\|) \qquad (7)$$

With respect to $m(x_0)$, $l(x_0)$, $r(x_0)$ and $l^{(x_j)}(x_0)$, $m^{(x_j)}(x_0)$, $r^{(x_j)}(x_0)$, $j = 1,2,3,\dots,p$ for the given kernel $K_h(.)$ and smoothing parameter h,

$$K_h(\|x_i - x_0\|) = \frac{K\left(\frac{\|x_i - x_0\|}{h}\right)}{h}, i = 1,2,3,\dots,n \qquad (8)$$

are a sequence of weights at x_0 whose role is to make the data that are close to x_0 contribute more in estimating the parameters at x_0 than those that are farther away with the adjustment of h. By solving this weighted least-squares problem, we can obtain not only the estimates of $m(x_0)$, $l(x_0)$ and $r(x_0)$ at x_0, but also those of their respective derivatives $l^{(x_j)}(x_0)$, $m^{(x_j)}(x_0)$, $r^{(x_j)}(x_0)$, $j = 1,2,3,\dots,p$. Since we mainly focus on estimating the underlying fuzzy non-parametric regression function

$$F(x) = \left(l(x), m(x), r(x)\right) \text{ at } x_0. \qquad (9)$$

It is natural to take the solutions of $m(x_0)$, $l(x_0)$ and $r(x_0)$ in equation (7), denoted, respectively by $\hat{m}(x_0)$, $\hat{l}(x_0)$ and $\hat{r}(x_0)$ as the estimates of the centre, the lower and the upper spread of F(x) at x_0. That is the estimate of F(x) at x_0:

$$\hat{F}(x) = \left(\hat{l}(x), \hat{m}(x), \hat{r}(x)\right)$$
$$= \left(\hat{m}(x_0) - \hat{\alpha}(x_0), \hat{m}((x_0), \hat{m}(x_0) - \hat{\beta}(x_0)\right) \qquad (10)$$

According to the principle of the weighted least-squares and by utilizing matrix notations, we can immediately obtain

$$\left(\hat{l}(x_0), \hat{l}^{(x_1)}(x_0), \dots, \hat{l}^{(x_p)}(x_0)\right)^T = \left(X^T(x_0)W(x_0;h)X(x_0)\right)^{-1} X^T(x_0)W(x_0;h)L_Y$$
$$\left(\hat{m}(x_0), \hat{m}^{(x_1)}(x_0), \dots, \hat{m}^{(x_p)}(x_0)\right)^T = \left(X^T(x_0)W(x_0;h)X(x_0)\right)^{-1} X^T(x_0)W(x_0;h)M_Y$$
$$\left(\hat{r}(x_0), \hat{r}^{(x_1)}(x_0), \dots, \hat{r}^{(x_p)}(x_0)\right)^T = \left(X^T(x_0)W(x_0;h)X(x_0)\right)^{-1} X^T(x_0)W(x_0;h)R_Y \quad (11)$$

Where

$$X(x_0) = \begin{bmatrix} 1 & x_{11} - x_{01} \dots x_{1p} - x_{0p} \\ 1 & x_{21} - x_{01} \dots x_{2p} - x_{0p} \\ \vdots & \vdots \\ 1 & x_{n1} - x_{01} \dots x_{np} - x_{0p} \end{bmatrix}$$

$$, L_Y = \begin{bmatrix} l_{y_1} \\ l_{y_2} \\ \vdots \\ l_{y_n} \end{bmatrix}, \quad M_Y = \begin{bmatrix} m_{y_1} \\ m_{y_2} \\ \vdots \\ m_{y_n} \end{bmatrix}, \quad R_Y = \begin{bmatrix} r_{y_1} \\ r_{y_2} \\ \vdots \\ r_{y_n} \end{bmatrix}$$

and

$$W(x_0;h) = \mathrm{diag}\left(K_h(\|x_1 - x_0\|)K_h(\|x_2 - x_0\|), \dots, K_h(\|x_n - x_0\|)\right). \qquad (12)$$

Thus, the estimated fuzzy regression function is

$$\hat{F}(x_0) = \left(\hat{l}(x_0), \hat{m}(x_0), \hat{r}(x_0)\right)_{T(R)}$$
$$= \left(e_1^T H(x_0;h)L_y, e_1^T H(x_0;h)M_y, e_1^T H(x_0;h)R_y\right) \qquad (13)$$

where

$$\mathrm{H}(\mathrm{x}_0;\mathrm{h}) = \left(\mathrm{X}^{\mathrm{T}}(\mathrm{x}_0)\mathrm{W}(\mathrm{x}_0;\mathrm{h})\mathrm{X}(\mathrm{x}_0)\right)^{-1}\mathrm{X}^{\mathrm{T}}(\mathrm{x}_0)\mathrm{W}(\mathrm{x}_0;\mathrm{h}) \tag{14}$$

and $e_1 = (1,0,\dots,0)^{\mathrm{T}}$, a (p+1)-dimensional vector with the first element being unity and the others being zero.

3.3 Fuzzy ridge nonparametric regression model

In most cases, due to multi co-linearity among independent variables, either the matrix $\left(\mathrm{X}^{\mathrm{T}}(\mathrm{x}_0)\mathrm{W}(\mathrm{x}_0;\mathrm{h})\mathrm{X}(\mathrm{x}_0)\right)$, is a singular matrix or it is very close to a singular matrix. In this paper, we use ridge regression to overcome this problem. Ridge regression gives computational efficiency in finding solutions of fuzzy regression models particularly for multi variable cases. In this case if we denote $\Theta, \mathrm{M}, \mathrm{W}, \widehat{\mathrm{W}}$ and N by

$$\Theta = (\theta_{11},\dots,\theta_{1n},\theta_{21},\dots,\theta_{2n},\theta_{31},\dots,\theta_{3n})^{\mathrm{T}},$$
$$\mathrm{M} = (\mathrm{M}_{\mathrm{Y}1},\dots,\mathrm{M}_{\mathrm{Y}n},\mathrm{M}_{\mathrm{Y}1}-\alpha_{\mathrm{Y}1},\dots,\mathrm{M}_{\mathrm{Y}n}-\alpha_{\mathrm{Y}n},\mathrm{M}_{\mathrm{Y}1} + \beta_{\mathrm{Y}1},\dots,\mathrm{M}_{\mathrm{Y}n}+\beta_{\mathrm{Y}n})^{\mathrm{T}},$$
$$\mathrm{W} = \mathrm{W}(\mathrm{x}_0;\mathrm{h}) = \mathrm{diag}\left(\mathrm{K}_{\mathrm{h}}(\|\mathrm{x}_1-\mathrm{x}_0\|),\mathrm{K}_{\mathrm{h}}(\|\mathrm{x}_2 - \mathrm{x}_0\|),\dots,\mathrm{K}_{\mathrm{h}}(\|\mathrm{x}_n-\mathrm{x}_0\|)\right), \tag{15}$$

$$\widehat{\mathrm{W}} = \begin{bmatrix} \mathrm{W} & 0 & 0 \\ 0 & \mathrm{W} & 0 \\ 0 & 0 & \mathrm{W} \end{bmatrix} \text{ and } \mathrm{N} = \begin{bmatrix} \mathrm{WQ} & 0 & 0 \\ 0 & \mathrm{WQ} & 0 \\ 0 & 0 & \mathrm{WQ} \end{bmatrix}$$

Where Q is a $n \times n$ matrix of $Q_{ij} = < x_i, x_j >$ and 0 is the $n \times n$ zero matrix. We have:

$$\Theta = 2\lambda(\mathrm{N}+\lambda\mathrm{I})^{-1}\widehat{\mathrm{W}}\mathrm{M}. \tag{16}$$

3.4 Selecting the kernel function and adjust smoothing parameter

After we use the above procedure to fit the fuzzy ridge non-parametric regression model (1), the regularization parameter λ, the Kernel $\mathrm{K}(.)$ and the smoothing parameter h in the weight $\mathrm{K}_{\mathrm{h}}(.)$ should be determined first. As discussed later the role of weights $\mathrm{K}(\|\mathrm{x}_i - \mathrm{x}_0\|)$, $i = 1,2,3,\dots,n$ is to make the data that are close to the given point x_0 contribute more to estimate $\hat{\mathrm{F}}(\mathrm{x})$ than those that are farther away. There are many types of kernel functions. In this study, we used Gaussian kernel

$$\mathrm{K}(\mathrm{x}) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\mathrm{x}^2}{2}\right). \tag{17}$$

Smoothing parameter h in the weight $\mathrm{K}_{\mathrm{h}}(.)$ is used to adjust the smoothness of the estimates $\hat{\mathrm{l}}(\mathrm{x}), \hat{\mathrm{m}}(\mathrm{x})$, and $\hat{\mathrm{r}}(\mathrm{x})$. Therefore, the proper selection of the smoothing parameter value is the important key point in the local smoothing techniques. There are a few approaches to selecting the optimal value of the smoothing parameter for the above local linear smoothing method, such as Bayesian and bootstrap, cross-validation and generalized cross-validation [20-22]. In this paper, a distance based on Diamond [18] to describe a fuzzified cross-validation procedure can be used and described as follows. Let

$$\hat{\mathrm{F}}_{(i)}(\mathrm{X}_i;\mathrm{h}) = \left(\hat{\mathrm{l}}_{(i)}(\mathrm{X}_i;\mathrm{h}),\hat{\mathrm{m}}_{(i)}(\mathrm{X}_i;\mathrm{h}),\hat{\mathrm{r}}_{(i)}(\mathrm{X}_i;\mathrm{h})\right)_{\mathrm{T(R)}} \tag{18}$$

be a predicted fuzzy ridge non-parametric regression function at input X_i computed by equation (13) with the smoothing parameter h. In this paper we use the following error evaluation criterion that we named it as CV (Cross-validation). In practice, to reach the optimal value of h closely depends on the degree of smoothness of the regression function we have to compute for a series of values of h to obtain h0. Afterward, we reach the most minimum value of proposed CV based on h0.

$$\mathrm{CV}(\mathrm{h}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{d}^2\left(\mathrm{Y}_i,\hat{\mathrm{F}}_i(\mathrm{x}_i,\mathrm{h})\right)\right)}$$
$$= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} \left(\mathrm{l}_{yi}-\hat{\mathrm{l}}_i(\mathrm{x}_i;\mathrm{h})\right)^2 + \\ \left(\mathrm{m}_{yi}-\hat{\mathrm{m}}_i(\mathrm{x}_i;\mathrm{h})\right)^2 + \left(\mathrm{r}_{yi}-\hat{\mathrm{r}}_i(\mathrm{x}_i;\mathrm{h})\right)^2\end{pmatrix}} \tag{19}$$

4. ReliefF-based algorithm

Beside the presented approach in section 3, in this paper we also introduce a simple but robust solution to achieve a good prediction. In the rest of the article we utilize the ReliefF weighting criteria as follows.

4.1 ReliefF basic concepts

ReliefF is an improved algorithm and more robust one, which can be used with incomplete and noisy data [23]. A key idea of the ReliefF algorithm is to estimate the quality of attributes (predictors) for input data and response a weighting vector for classification or regression with K nearest neighbors. Weight vector, is consisting of attribute weights ranging from [-1..1] with large positive weights assigned to important attributes. Attribute weights computed by ReliefF usually depend on accurate selection

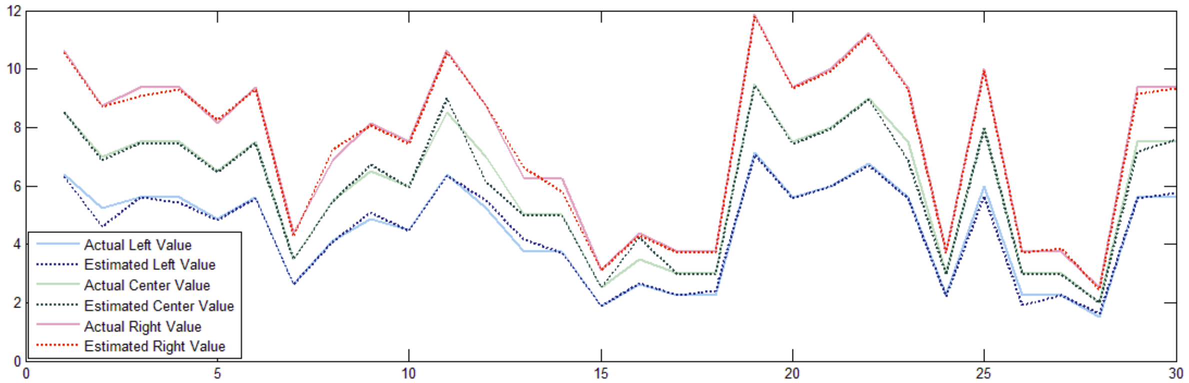

Fig. 2 Prediction accuracy of testing data in FNPreg, RMSE= 0.0031, Accuracy= 0.994

of K nearest neighbors. User-defined parameter k controls the locality of the estimates. In most proposes, this parameter can start with 10, so in the presented algorithm it can be safely set to 10 too.

4.2 Proposed Algorithm

In algorithm 1 the pseudo code of the presented approach has been provided. Vector W is consisted of ReliefF weighting factors (line 4) and utilize with a metric element to make a metric vector called S (lines 5, 6, 7). Finally the index of minimum value in vector S is mapped to a class position in training set and its value returned as outcome (lines 8, 9, 10).

Algorithm 1 proposed ReliefF-based Algorithm

```
1. set all weightsW[A] = 0.0;
2. set C1 = count(TestingSet);
3. set C2 = count(TrainingSet);
/* Estimate attributes weight and make weighting vector W */
4. W[A] = ReliefF(TrainingSet, k)
5. for i = 1 to C1 do begin
6.     for j = 1 to C2 do begin
7.         $Set[j] = \sum(W[A_i] \times |TestSet\_A(i) - TrainingSet\_A(j)|)$;
       /* find the minimum value in Set with its position */
8. value = minimum(Set);
9. position = indexvalue;
10. return(TrainingSet_Class[position]);
11. end;
```

5. Case study

This study employs actual training and testing datasets exported to CSV format with totally 108 records from students in Damghan University during an academic year from September to June in machine learning course using the Moodle platform [1]. This platform can store some specific tasks carried out by the students during an academic year, just before the Final Examinations. In order to collect information, each user in the system is assigned an identifier and every time he/she logs onto the framework, all movements within a particular course are stored with respect to his/her access to content and tools (e.g. calendar, discussion forum, email archive, chat, quizzing, and assignments) [24]. In our work, both the information about four quizzes that taken from students along the semester as well as final marks obtained in this course, are considered.

6. Experimentation and results

This section discusses the experimental results. First section provides the results of the traditional algorithms as shown in table 1. And second section provides the performance of our proposal that is provided in table 2.

6.1 Experimental results of traditional algorithms

The purpose of the experimentation is to show that proposed approaches improve the efficiency and effectiveness of the classical representation, in predict student’s grades. Thus, first a comparative study is carried out between the most applicable algorithms and then the presented proposal is evaluated for solving the same

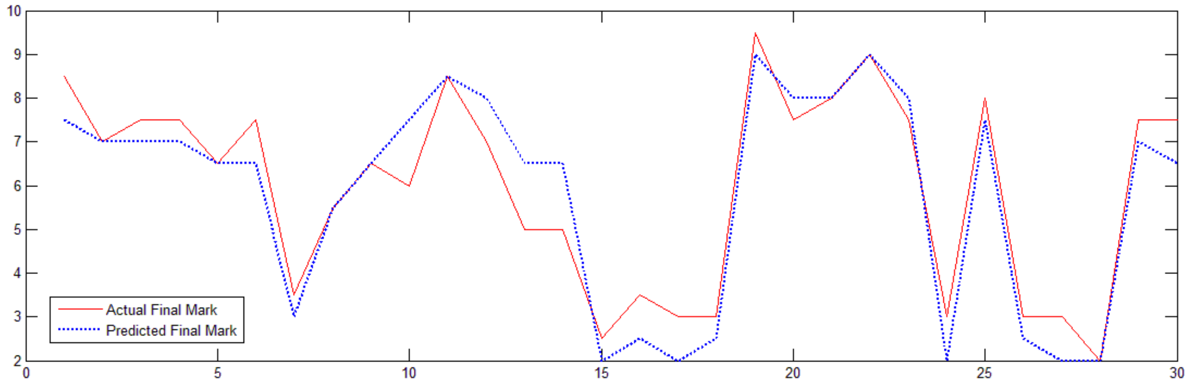

Fig. 3 Prediction accuracy of testing data in ReliefF-based algorithm, RMSE= 0.6236, Accuracy= 0.9075

problem. In order to hold all experiments in the same conditions, each of them using 10-fold stratified cross-validation, and they are reported in this section. To compare the proposed approach we consider some popular paradigms that have shown good results in other applications.

Table 1 report Root Mean Squared Error and accuracy using R2 criterion for the measurements for all algorithms employed in this study.

Table 1 Experimental results of traditional methods

Algorithms	Accuracy	RMSE
Algorithms based on rules		
M5Rules	0.7066	1.4709
Bagging	0.7049	1.4481
DecisionTable	0.5662	1.6357
ConjunctiveRule	0.3974	1.7529
Algorithms based on decision tree		
M5P	0.7066	1.4709
RepTree	0.4694	1.6965
DecisionStump	0.3046	1.9544
Algorithms based on regression		
PaceRegression	0.7200	1.4635
LinearRegression	0.7066	1.4709
SMOreg	0.6865	1.3402
SimpleLinearRegression	0.6398	1.6088
LeastMedianSquared	0.6383	1.9360
IsotonicRegression	0.5662	1.6140
AdditiveRegression	0.5029	1.6572
Algorithm based on Neural Network		
GuassianProcess	0.7628	1.4188
MultilayerPerceptron	0.6961	1.7692
RBFNetwork	0.5722	1.6211

6.2 Results obtained using our proposals

In this section we discuss the performance achieved by our proposal. Number of records is chosen to be 30 percent randomly chosen as testing set and number of running algorithms are about over 25 times. Results of two proposed algorithms are shown as follows:

- Fuzzy nonparametric regression

  CV outcomes for some smoothing argument (h) and regularization parameter (λ) is adjusted and best value is computed around $h = 0.128$ and $\lambda = 0.0002$. Figure 2 shows the prediction results for randomly 30 percent samples as fuzzy triangular final grades. At a glance, the accuracy of the prediction is clearly seen and it can be shown that the presented approach returns values much nearest to the actual ones beside traditional presented algorithms.

- ReliefF-based algorithm

  Figure 4 shows the prediction diagram for 30 percent randomly chooses. And figure 5 shows the accuracy of the presented algorithm using R2 criterion.

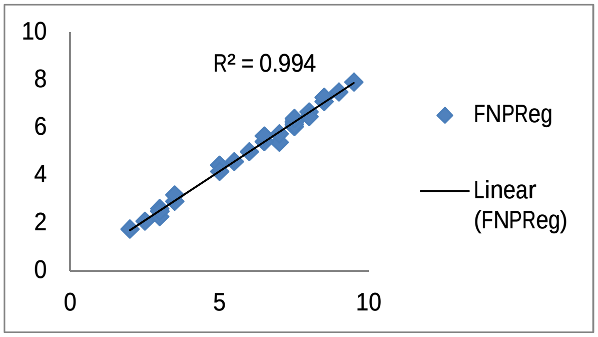

Fig. 4 Prediction accuracy of fuzzy non-parametric regression (FNPreg)

In the following, table 2 shows the accuracy and root mean squared error (RMSE) provided by each proposed algorithm of testing datasets. The values indicates that the proposed models completely overcome all traditional algorithms and make much more accurate prediction.

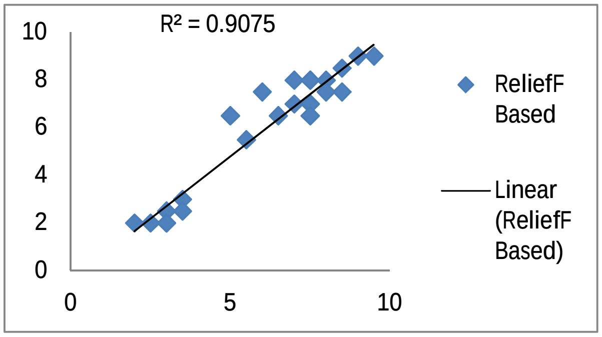

Fig. 5 Prediction accuracy of ReliefF-based algorithm

Table 2 Results obtained using our proposals

Algorithms	Accuracy	RMSE
Fuzzy non-parametric regression	0.9940	0.0031
ReliefF-based	0.9075	0.6236

7. Conclusion and future work

To achieve more accuracy in prediction process in educational environments, two new proposals based on statistical approach were presented. First we introduced a novel approach based on fuzzy non-parametric regression by integrating ridge-type regularization in the Lagrangian dual space and using Gaussian kernel as well as smoothing parameter, all together to fit the presented model. Second we integrate ReliefF weighting algorithm as a weighting vector in the proposed algorithm. Computational experiments show that when the problem is regarded as fuzzy non-parametric regression or ReliefF-based algorithm, performance is significantly better and the weakness of all other traditional proposed algorithms is overcome.

Although the results are so interesting, there are still quite a few considerations that could surely add even more value to results obtained. As considered later this study is concentrated only on crisp inputs such as online students' quizzes marks, to predict final grades, so it would be interesting to design a method to apply desirable categorical variables together. Another interesting work is implementing a method with integrating ridge-type regularization in fuzzy nonlinear regression, in which it can be so more value to the work.

References

[1] Zafra, A., Ventura, S. (2012) Multi-instance genetic programming for predicting student performance in web based educational environments. Applied Soft Computing. 12(8): 2693–2706.

[2] Mostow, J., Beck. ,J. (2006) Some useful tactics to modify, map and mine data from intelligent tutors. Natural Language Engineering, 12(2): 195-208.

[3] Minnaert, A., Janssen, P. (1999) The additive effect of regulatory activities on top of intelligence in relation to academic performance in higher education. Learning and Instruction, 9, pp. 77–91.

[4] Busato, V., Prins, F., Elshout, J., Hamaker, C. (2000) Intellectual ability, learning style, personality, achievement motivation and academic success of psychology students in higher education. Personality and Individual Differences, 29, pp. 1057–1068.

[5] Delavari, N., Phon-amnuaisuk, S., Beikzadeh, M.R. (2008) Data Mining Application in Higher Learning Institutions. Informatics in Education. 1(7): 31–54.

[6] Luan, J., Zhao, C.M. ,Hayek, J. (2004) Use data mining techniques to develop institutional typologies for NSSE. National Survey of Student Engagement.

[7] Zafra, A., Ventura, S. (2010) Web Usage Mining for Improving Students Performance in Learning Management Systems. 6098, pp. 439-449.

[8] Chanchary, F.H., Haque, I., Khalid, M.S. (2008) Web Usage Mining to Evaluate the Transfer of Learning in a Web-based Learning Environment. Knowledge Discovery and Data Mining. WKDD 2008. pp. 249–253.

[9] Fausett, L., Elwasif, W. (1994) Predicting Performance from Test Scores using Backpropagation and Counterpropagation. IEEE Congress on Computational Intelligence, pp. 3398–3402.

[10] Martnínez, D. (2001) Predicting student outcomes using discriminant function analysis. In: Annual Meeting of the Research and Planning Group, California, USA. pp. 163–173.

[11] Minaei-Bidgoli, B., Punch, W. (2003) Using genetic algorithms for data mining optimization in an educational

web-based system. Genetic and Evolutionary Computation, 2, pp. 2252–2263.

[12] Kotsiantis, S., Pintelas, P. (2005) Predicting students marks in Hellenic open university. In ICALT'05: The fifth international conference on advanced learning technologies. Kaohsiung, Taiwan, pp. 664–668.

[13] Romero, C., Espejo, P.G., Zafra, A., Romero, J.R., Ventura, S. (2010) Web Usage Mining for Predicting Final Marks of Students That Use Moodle Courses. Computer Applications in Engineering Education. 1(21): pp. 1–12.

[14] Zafra, A., Romero, C., Ventura, S. (2011) Multiple instance learning for classifying students in learning management systems. Expert Systems with Applications. 12(38): pp. 15020–15031.

[15] Romero, C., Ventura, S. (2010) Educational data mining: A review of the state-of-the-art. IEEE Transaction on Systems, Man and Cybernetics Applications and Reviews. 40(6): 610–618.

[16] H. Tanaka, H. Ishibuchi, Identification of possibilistic linear systems by quadratic membership functions of fuzzy parameters (1991) Fuzzy Sets and Systems 2(41): pp. 145-160.

[17] C.B. Cheng, E.S. Lee (2001) Fuzzy regression with radial basis function networks. Fuzzy Sets and Systems 2(16): pp. 291–301.

[18] P. Diamond, Fuzzy least squares (1988). Inform. Sci. 3(46): pp. 141–157.

[19] Kim B, Bishu R. R. (1998) Evaluation of fuzzy linear regression models by comparing membership functions. Fuzzy sets and systems, 100, 343 – 352.

[20] J. Fan, I. Gijbels (1996) Local Polynomial Modeling and Its Applications, Chapman & Hall, London.

[21] J. D. Hart (1997) Nonparametric Smoothing and Lack-of-fit Tests, Springer-Verlag, New York.

[22] D. H. Hong, C. Hwang, C. Ahn (2004) Ridge estimation for regression models with crisp inputs and Gaussian fuzzy output. Fuzzy Sets and Systems, 142, 307-319.

[23] Robnik M.S., Kononenko I. (2003) Theoretical and Empirical Analysis of ReliefF and RReliefF. Machine Learning Journal. pp. 23-69.

[24] Zafra A., Romero C., Ventura S. (2011) Multiple instance learning for classifying students in learning management systems. Expert Systems with Applications. 12(38), pp. 15020–15031

[25] S. A. Diwani and A. Sam, "Data Mining Awareness and Readiness in Healthcare Sector : a case of Tanzania," ACSIJ Adv. Comput. Sci. an Int. J., vol. 3, no. 1, pp. 37–43, 2014.