**Journal of Industrial Strategic Management**
**Nov. 2016, Vol. 1, No. 3, 1-10** **Islamic Azad University, Firoozkooh Branch**

# Fuzzy Data Envelopment Analysis for Classification of Streaming Data

**Alireza Alinezhad**
Associate Professor, Faculty of industrial and mechanical engineering, Qazvin branch, Islamic Azad University, Qazvin, Iran
E-mail address: alinezhad_ir@yahoo.com,

**Amineh Tohidi**
Msc. Student, Faculty of industrial and mechanical engineering, Qazvin branch, Islamic Azad University, Qazvin, Iran
E-mail: amineh_tohidi@yahoo.com

**Mohammad Amin Adibi**
Assistant Professor, Faculty of industrial and mechanical engineering, Qazvin branch, Islamic Azad University, Qazvin, Iran
E-mail: adibi@qiau.ac.ir

## Abstract

*The classification of fuzzy uncertain data is considered one of the most challenging issues in data analysis. In spite of the significance of fuzzy data in mathematical programming, the development of the analytical methods of fuzzy data is slow. Therefore, the current study proposes a new fuzzy data classification method based on fuzzy data envelopment analysis (DEA) which can handle streaming data.The new method is tested by simulated data and the results indicate its effectiveness in facing uncertain data and variable conditions.*

**Keywords:** Fuzzy Data Envelopment Analysis, Mathematical Programming; Classification; Fuzzy Streaming Data

## Introduction

Classification is assigning a class or category to a data (position) based on a predetermined function or model. Such a model is obtained through comparing the class of so-called training data series using various methods such as decision tree, Artificial Neural Network (ANN), Support Vector Machine (SVM), logistic regression, etc. (Gazanfari, et al. 1387). The model determines whether data belongs to a particular class. The classification is one of the most common issues in data analysis for predicting the class of objects and situations, identifying abnormalities or factors affecting certain phenomenon, etc.

One of the challenges of classification is to create models for classifying uncertain fuzzy data. In many cases the collected data are not certain for various reasons and sometimes the data are not collected toggether but are observed over time. The latter is often referred as streaming data (Mena-Torres & Aguilar-Ruiz 2014). In case of the classification of the streaming data, classification is even more complicated because the fuzzy uncertain aspects of the data should be updated in the system as they change over the time.

In such conditions, Data Envelopment Analysis (DEA) is the most appropriate method for classification. Although, DEA is used to empirically measure productive efficiency of decision making units, it can be applied in other applications such as classification of data (Yan, & Wei 2011; Pendharkar 2011). In this case, each data is considered a DMU where data characteristics are inputs and classes are outputs. DNA method can be used for classification of fuzzy data through a linear programming (Pendharkar 2012; Taneja et al 2016). Another advantage of using DEA for classification is ease of the solution modification in linear programming. Thus, classification model can be updated running sensitivity analysis.

According to the researchers, this special advantage of DEA has not received much attention. In fact, such an advantage can be used to construct a framework to handle streaming data with concept drift, which is the main contribution of this paper. Considering that the computational time required for solving linear programming model greatly increases when the number of variables and constraints increases, the use of a mechanism to control the aspects of a problem is required over time. This mechanism should be able to provide the suitable ground for optimization through controlling entry and exit of effective data on the efficiency frontier over time: this is also studied in the paper.

The paper is structured as follows: Section 2 presents data classification using DEA; Section 3 provides fuzzy data classification; Section 4 discusses the proposed method; Section 5 presents model testing; and Section 6 sums up the results.

## Data Classification Using DEA

Our goal is to classify the data by identifying a border (model). If any data is considered a DMU, so that the values of the characteristics of each data are inputs of DMU and 1 is its output, data known as frontier point in DEA can be used to illustrate the range (or border) of the category. Then, these ranges can be used to predict the category or class of new data. This means if

$$x_i = (x_{i1}, x_{i2}, \dots, x_{ij}, \dots, x_{im})$$

where $i = 1,2,\dots,n$ is in one category, the range of category can be determined through solving a set of linear programming problems in the form of a

DEA problem to identify the border areas, as shown in Equation 1

$$Minimize \theta^{t}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \xi^{t} x_{tj} \leq 0, \quad j = 1, \ldots, m \qquad (1)$$
$$\sum_{i=1}^{n} \lambda_i = 1$$
$$\lambda_i \geq 0, \; i = 1, \ldots, n.$$

As an example, we provide the data in Table 1 which have two classes 1 and 2. The data was obtained from the study of Pendharkar and Troutt 2014.

Equation 2 and Equation 3 were applied to solve two series of DEA problems and to identify the ranges of two classes. Figure 2 presents the range obtained for two classes based on the border areas. In addition, dependent variable of θ is given in Tables 2 and 3 which was used to develop an LP models related to the first and second classes based on which the identification of border areas was performed

Table 1. Example Data

| 1st feature | 2nd feature | Class | 1st feature | 2nd feature | Class |
|---|---|---|---|---|---|
| 640 | 6.02 | 1 | 310 | 4.7 | 2 |
| 550 | 6.09 | 1 | 350 | 4.5 | 2 |
| 510 | 5.67 | 1 | 400 | 4.7 | 2 |
| 420 | 5.54 | 1 | 370 | 4.8 | 2 |
| 560 | 6.75 | 1 | 450 | 4.7 | 2 |
| 550 | 6.60 | 1 | 500 | 4.5 | 2 |
| 580 | 5.87 | 1 | 520 | 4.6 | 2 |
| 420 | 6.20 | 1 | 550 | 4.3 | 2 |
| 450 | 6.77 | 1 | 570 | 4.5 | 2 |
| 520 | 5.67 | 1 | 450 | 4.9 | 2 |
| 440 | 5.33 | 1 | 320 | 4.6 | 2 |
| 480 | 5.96 | 1 | 400 | 4.6 | 2 |
| 520 | 6.13 | 1 | 310 | 5.1 | 2 |
| 570 | 6.26 | 1 | | | |
| 400 | 5.95 | 1 | | | |
| 580 | 5.2 | 1 | | | |

$$Minimize \theta^{t}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \theta^{t} x_{tj} \leq 0, \quad j = 1,2 \qquad (2)$$
$$\sum_{i=1}^{16} \lambda_i = 1$$
$$\lambda_i \geq 0, i = 1, \ldots, 16.$$

$$Maximize \theta^{t}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i x_{ij} - \theta^{t} x_{tj} \geq 0, \quad j = 1,2 \qquad (3)$$
$$\sum_{i=1}^{13} \lambda_i = 1$$
$$\lambda_i \geq 0, i = 1, \ldots, 13.$$

Table 2. Value of θ related to the data of class 1 in LP model (based on the data of Table 1)

| Data (Class1) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | 0.87 | 0.87 | 0.93 | 1.00 | 0.79 | 0.81 | 0.90 | 0.96 | 0.89 | 0.93 | 1.00 | 0.90 | 0.87 | 0.85 | 1.00 | 1.00 |

Table 3. Value of θ related to the data of class 2 in LP model (based on the data of Table 1)

| Data (Class:2) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | 108 | 1.11 | 1.05 | 1.04 | 1.03 | 1.04 | 1.01 | 1.04 | 1.00 | 1.00 | 1.10 | 1.07 | 1.00 |

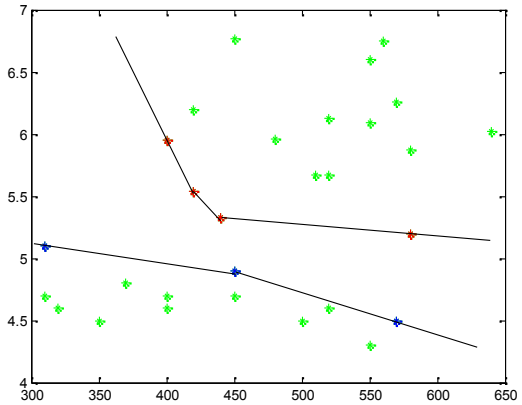


Figure 1. The border of classes 1 and 2 based on frontier points and the data of Table 1

Classification of Uncertain Fuzzy Data Using Data Envelopment Analysis

Assuming that the value of jth ($j = 1,2,\dots,m$) characteristic related to ith data ($i = 1,2,\dots,n$) is a trapezoidal fuzzy number in the form of $\tilde{x}_{ij} = (a_{ij}, b_{ij}, c_{ij}, d_{ij})$, the pattern of linear programming model appropriate to a DEA problem will be in the form of Equation 4. The membership function of a trapezoidal fuzzy number is presented in figure 2.

$$minimize\ \theta^{t}$$

Subject to:

$$\sum_{i=1}^{n} \lambda_i \tilde{x}_{ji} \leq \theta^{t} \tilde{x}_{jt}, \quad j = 1, \ldots, m$$
$$\sum_{i=1}^{n} \lambda_i \tilde{y}_{ri} \geq \tilde{y}_{rt}, \quad r = 1, \ldots, s \qquad (4)$$
$$\sum_{i=1}^{n} \lambda_i = 1$$
$$\lambda_i \geq 0, \quad i = 1, \ldots, n$$

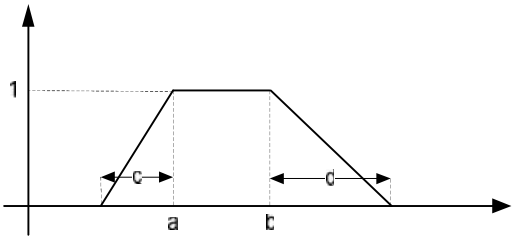


Figure 2. Membership function of a trapezoidalfuzzy number $\tilde{x} = (a, b, c, d)$

León, et al (2003) and Lampe and Hilgers (2015) provide an Equation 4 to receive values of θ and develop an LP model presented in Equation 5 where $x^L = a$, $x^R = b$, $\alpha^L = c$ , $\alpha^R = d$ and s is the output of each DMU.

Table 4. Fuzzy Numbers Related to the Example Fuzzy Data

| Data | x1_a | x1_b | x1_c | x1_d | x2_a | x2_b | x2_c | x2_d | Class |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 3.25 | 4.00 | 0.50 | 0.50 | 3.50 | 4.25 | 0.50 | 0.50 | 1 |
| 2 | 3.00 | 4.00 | 0.50 | 0.50 | 4.75 | 5.25 | 0.50 | 0.50 | 1 |
| 3 | 3.00 | 4.00 | 0.50 | 0.50 | 5.75 | 6.25 | 0.50 | 0.50 | 1 |
| 4 | 2.00 | 3.00 | 0.50 | 0.50 | 5.75 | 6.25 | 0.50 | 0.50 | 1 |
| 5 | 4.00 | 5.00 | 0.50 | 0.50 | 4.75 | 5.25 | 0.50 | 0.50 | 1 |
| 6 | 5.00 | 6.00 | 0.50 | 0.50 | 3.00 | 3.50 | 0.50 | 0.50 | 1 |
| 7 | 5.00 | 6.00 | 0.50 | 0.50 | 6.00 | 6.50 | 0.50 | 0.50 | 1 |
| 8 | 4.00 | 5.00 | 0.50 | 0.50 | 3.75 | 4.25 | 0.50 | 0.50 | 1 |
| 9 | 4.96 | 5.96 | 0.50 | 0.50 | 5.31 | 5.81 | 0.50 | 0.50 | 1 |
| 10 | 5.26 | 6.26 | 0.50 | 0.50 | 5.38 | 5.88 | 0.50 | 0.50 | 1 |
| 11 | 1.50 | 2.50 | 0.50 | 0.50 | 0.50 | 1.00 | 0.50 | 0.50 | 2 |
| 12 | 1.25 | 2.25 | 0.50 | 0.50 | 1.75 | 2.25 | 0.50 | 0.50 | 2 |
| 13 | 1.25 | 2.25 | 0.50 | 0.50 | 2.75 | 3.25 | 0.50 | 0.50 | 2 |
| 14 | 0.25 | 1.25 | 0.50 | 0.50 | 2.75 | 3.25 | 0.50 | 0.50 | 2 |
| 15 | 2.25 | 3.25 | 0.50 | 0.50 | 1.75 | 2.25 | 0.50 | 0.50 | 2 |
| 16 | 3.25 | 4.25 | 0.50 | 0.50 | 2.00 | 2.50 | 0.50 | 0.50 | 2 |
| 17 | 3.25 | 4.25 | 0.50 | 0.50 | 2.00 | 2.50 | 0.50 | 0.50 | 2 |
| 18 | 2.25 | 3.25 | 0.50 | 0.50 | 0.75 | 1.25 | 0.50 | 0.50 | 2 |
| 19 | 3.21 | 4.21 | 0.50 | 0.50 | 2.31 | 2.81 | 0.50 | 0.50 | 2 |
| 20 | 3.51 | 4.51 | 0.50 | 0.50 | 2.38 | 2.88 | 0.50 | 0.50 | 2 |

As shown in Table 4, 20 trapezoidal fuzzy data belong to two classes 1 and 2. For these data, two classes have been identified using the model. The results are shown in Figure 3. It should be noted that maximum-average of membership degree, $(a+b)/2$ is used to obtain fuzzy data of their corresponding equivalents

$$
\begin{aligned}
&\textit{minimize } \theta^t \\
&\textit{Subject to:} \\
&\sum_{i=1}^{n} \lambda_i x_{ji}^L \le \theta^t x_{jt}^L, \quad j = 1, \dots, m \\
&\sum_{i=1}^{n} \lambda_i x_{ji}^R \le \theta^t x_{jt}^R, \quad j = 1, \dots, m \\
&\sum_{i=1}^{n} \lambda_i x_{ji}^L - \sum_{i=1}^{n} \lambda_i \propto_{ji}^L \le \theta^t x_{jt}^L - \theta^t \propto_{jt}^L, j = 1, \dots, m \\
&\sum_{i=1}^{n} \lambda_i x_{ji}^R + \sum_{i=1}^{n} \lambda_i \propto_{ji}^R \le \theta^t x_{jt}^R - \theta^t x_{jt}^R, \quad J = 1, \dots, m \\
&\sum_{i=1}^{n} \lambda_i y_{ri}^L \ge y_{rt}^L, \quad r = 1, \dots, s \\
&\sum_{i=1}^{n} \lambda_i y_{ri}^R \ge y_{rt}^R, \quad r = 1, \dots, s, \\
&\sum_{i=1}^{n} \lambda_i y_{ri}^L - \sum_{i=1}^{n} \lambda_i \beta_{ri}^L \le y_{rt}^L - \beta_{rt}^L, \quad r = 1, \dots, s \\
&\sum_{i=1}^{n} \lambda_i y_{ri}^R + \sum_{i=1}^{n} \lambda_i \beta_{ri}^R \ge y_{rt}^R - \beta_{rt}^R, \quad r - 1, \dots, s \\
&\sum_{i=1}^{n} \lambda_i = 1, \\
&\lambda_i \ge 0, \quad i = 1, \dots, n
\end{aligned}
\tag{5}
$$

Table 5. Value of $\theta$ Related to the Data of Class 1 in Fuzzy LP Model (based on the data of Table 4)

| Data (Class 1) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | 1.00 | 0.94 | 0.87 | 1.00 | 0.82 | 1.00 | 0.69 | 0.95 | 0.74 | 0.72 |

Table 6. Value of $\theta$ Related to the data of Class 2 in Fuzzy LP Model (based on the data of Table 4)

| Data (Class 2) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\theta$ | 1.67 | 1.31 | 1.00 | 1.00 | 1.25 | 1.06 | 1.06 | 1.34 | 1.03 | 1.00 |

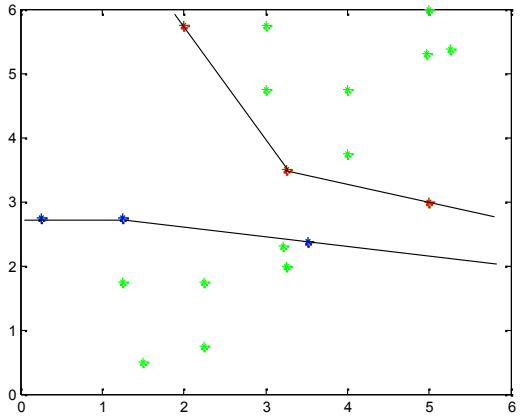


Figure 3. The border of classes 1 and 2 based on frontier points and the data of Table 4

## Classification of Fuzzy Streaming Data Using DEA

The new framework, which is used to classify fuzzy data using data envelopment analysis, is illustrated in Figure 4. As it is shown in the figure, first, DEA problems are solved to identify boundary areas based on training fuzzy data, then, the label of each data is specified. Borders obtained in each DEA problem are used to determine benchmarking data for that category. Before starting the classification of the data stream, the variable D, which is equal to zero, is defined. The variable is applied in the process of classification of the streaming data to maintain the distance of data, which do not confirm the model of the identified classes. In addition, $S_{new}$ is defined and used to collect the data, which do not confirm the model of the recognized classes. At the beginning this set is empty.

In the classification of the streaming data, if new data y is assigned to one of the classes according to the identified current ranges, we determine the label of the new data in accordance with the classification to which they belong.

To calculate the distance between two points equivalent to two trapezoidal fuzzy numbers, the equation 6 is used otherwise the label of data equals to the label of the nearest vector $y_n$ (León et al 2003). Meanwhile, the distance between $y$ and $y_n$ is added to D and $y$ is added to $S_{new}$. If D is lower than the predetermined threshold limit of $\omega$, the range of classes is determined for the next data based on the current border areas. Otherwise, it can be deduced, that the current limits are not significantly able to cover all the data, i.e. the system has changed over the time and the change in the system has been manifested in the form of a change in the behavioural pattern of data. Therefore, the modification of the previous ranges is required.

$$d(\tilde{x}_1, \tilde{x}_2) = \sqrt{\frac{1}{6}\left[\left((a_1 - c_1) - (a_2 - c_2)\right)^2 + 2(a_1 - a_2)^2 + 2(b_1 - b_2)^2 + ((b_1 + d_1) - (b_2 + d_2))^2\right]} \quad (6)$$

To modify ranges, given that the basis of areas is to solve DEA problems, new DEA problems should be solved based on a new dataset. The set of the new data is collected by adding the members of $S_{new}$ to the set of the current training data and removing as many previous members as added. This method, which is called Windowing Technique, keeps the number of the members of the training dataset constant and enables to create new boundaries for the correct classification based on the status of the system. After these procedures and modifications, new boundary points of the training dataset are obtained and the ranges of new classes can be identified. At this stage, D value is equal to zero and $S_{new}$ becomes empty. Then, the next new data can be classified with updated ranges. These measures will maintain until there is streaming data.

## Testing the Proposed Classification Framework

To generate initial training data as well as data that fit the proposed model, the study conducted by Yazdi et al. (2009) is considered. However, in their study, only

classification of fuzzy data in static mode has been considered. Hence, suitable streaming data should be created. Consequently, first, the initial 40 (two-dimensional) training data were generated according to the tables 7 and 8. Then, streaming data were generated according to the methods outlined in Tables 9 and 10. As a result, 180 fuzzy data were randomly generated in this phase, while in total 40 initial training data and 220 streaming data were generated. Figure 4 illustrates the proposed method.

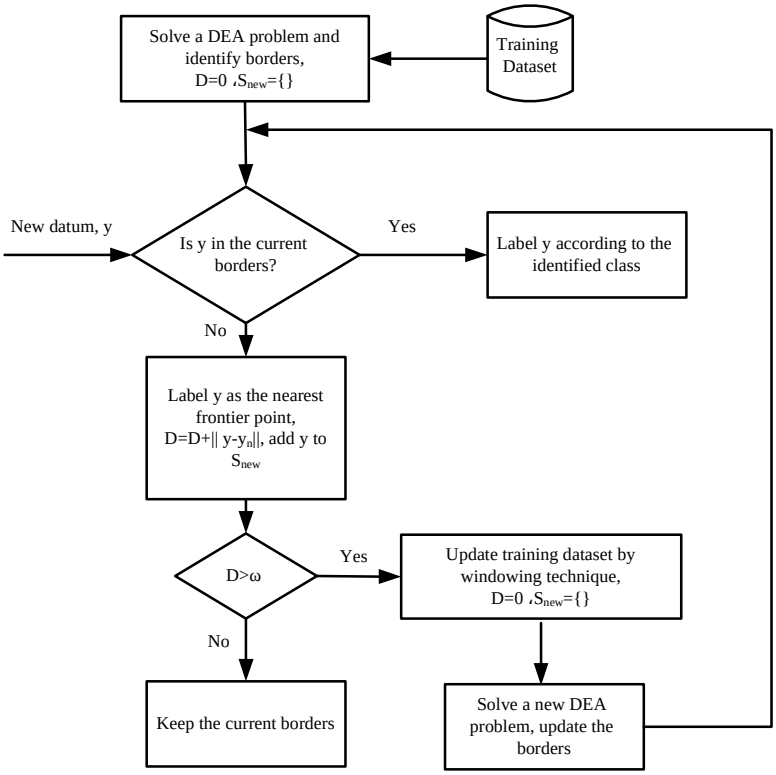


Figure 4. The proposed method for classification of streaming fuzzy data

Table 7. Way of generating fuzzy data for the class 1 to use in the initial stage of the test

| | |
|---|---|
| 1st feature | $a \sim N(7,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |
| 2nd feature | $a \sim N(9,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |

Table 8. Way of genrating fuzzy data for the class 2 to use in the initial stage of the test

| | |
|---|---|
| 1st feature | $a \sim N(3,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |
| 2nd feature | $a \sim N(3,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |

Table 9. Way of generating fuzzy data for the class 1 to use in the final stage of the test

| | |
|---|---|
| 1st feature | $a \sim N(8,1.5), b' \sim U(0,1), b = a + b',\ c,d \sim U(0.2,0.7)$ |
| 2nd feature | $a \sim N(9,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |

Table 10. Way of generating fuzzy data for the class 2 to use in the final stage of the test

| | |
|---|---|
| 1st feature | $a \sim N(2,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |
| 2nd feature | $a \sim N(4,1.5),\ b' \sim U(0,1),\ b = a + b',\ c,d \sim U(0.2,0.7)$ |

After running the new method of classification of the fuzzy data stream using DEA with MATLAB software, F1-Score, which is defined as Equation 7, is used to assess the proposed method. The precision of the model is based on the proportion of the data of the considered category and recall is a ratio of the data that properly belong to a particular category. Table 11 illustrates the results of the application of the proposed method. It is worth mentioning that YALMIP software implemented in MATLAB has been used to solve linear programming problem.

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (7)$$

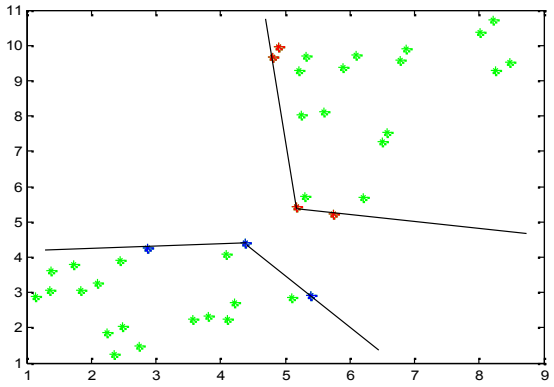


Figure 5. The border of classes 1 and 2 based on frontier points related to initial training dataset

Table 11. The value of Recall, Precision and F1-measure for test fuzzy data stream

| | Precision | Recall | F1-Score |
|---|---|---|---|
| Class 1 | 0.92 | 1.00 | 0.96 |
| Class 2 | 1.00 | 0.78 | 0.87 |

**Results**

This study developed a model of the data classification in the form of a Data Envelopment Analysis (DEA) problem based on the trapezoidal fuzzy uncertain data. The proposed model is a new classification model of fuzzy streaming data, which enables to update the border of classes by updating data over time. The new proposed methods were used on the simulated data from the literature and as a result, F1-Score for the prediction of classification reached 0.92. The proposed method only needs an input parameter $\omega$ to be determined by the user.

**References**

1. Chen, T. Y., Ku, T. C., & Tsui, C. W., 2008. Determining attribute importance based on triangular and trapezoidal fuzzy numbers in (z) fuzzy measures. In The 19th International Conference on Multiple Criteria Decision Making (pp. 75-76).
2. Gazanfari, M., Alizadeh, S., & Teimourpour, B., 1387. Data Mining and Knowledge Discovery. Iran University of Science and Technology. (In Persian).
3. Lampe, H.W. and Hilgers, D., 2015. Trajectories of efficiency measurement: A bibliometric analysis

of DEA and SFA. European Journal of Operational Research, 240(1), pp.1-21.

4. León, T., Liern, V., Ruiz, J. L., & Sirvent, I., 2003. A fuzzy mathematical programming approach to the assessment of efficiency with DEA models. Fuzzy sets and systems, 139(2), 407-419.
5. Mena-Torres, D., & Aguilar-Ruiz, J. S., 2014. A similarity-based approach for data stream classification. Expert Systems with Applications, 41(9), 4224-4234.
6. Pendharkar, P., 2012. Fuzzy classification using the data envelopment analysis. Knowledge-Based Systems, 31, pp.183-192.
7. Pendharkar, P.C. and Troutt, M.D., 2014. Interactive classification using data envelopment analysis. Annals of Operations Research, 214(1), pp.125-141.
8. Pendharkar, P.C., 2011. A hybrid radial basis function and data envelopment analysis neural network for classification. Computers & Operations Research, 38(1), pp.256-266.
9. Taneja, S., Suri, B., Narwal, H., Jain, A., Kathuria, A. and Gupta, S., 2016, January. A new approach for dataf classification using Fuzzy logic. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) (pp. 22-27). IEEE.
10. Yan, H., & Wei, Q., 2011. Data envelopment analysis classification machine. Information Sciences, 181(22), 5029-5041.
11. Yazdi, H. S., & Vhedian, A., 2009. Fuzzy Bayesian classification of LR Fuzzy numbers. IACSIT International Journal of Computer Theory and Engineering, 1(5).