American Journal of Applied Mathematics
2015; 3(2): 75-80
Published online March 31, 2015 (http://www.sciencepublishinggroup.com/j/ajam)
doi: 10.11648/j.ajam.20150302.17
ISSN: 2330-0043 (Print); ISSN: 2330-006X (Online)

Clustering Problem with Fuzzy Data: Empirical Study for Financial Distress Firms

Slah Benyoussef[1, 2]

[1]Airport Rd, Al-Imam Muhammad Ibn Saud Islamic University, Riyadh 11432, Arabie Saoudite
[2]Faculté des Sciences Economiques et de Gestion de Sfax, route Aéroport km 4, BPN°1088, 3018 Sfax, Tunisie

Email address:
benyoussef_slah@yahoo.fr

To cite this article:
Slah Benyoussef. Clustering Problem with Fuzzy Data: Empirical Study for Financial Distress Firms. *American Journal of Applied Mathematics*. Vol. 3, No. 2, 2015, pp. 75-80. doi: 10.11648/j.ajam.20150302.17

Abstract: In many real applications, the data of classification problems cannot be precisely measured. However, in an increasingly complex environment, these variables can be imprecise, qualitative or linguistic. In such a case, fuzzy set theory seems to be the convenient tool to fill this insufficiency. Thus, we proposed a new approach, based on the ranking function, which consists in solving the classification problems via fuzzy linear programming model. This approach has been applied for the financial distress firms. The obtained results are satisfactory in terms of correctly classified rates

Keywords: Bankrupcy firms, Classification problems, Fuzzy logic, Linear programming, Ranking function

1. Introduction

The bankruptcy problem has become more and more important as the competition between financial institutions has come to a totally conflicting stage. More and more companies are seeking better strategies through the help of credit scoring models and hence discriminant analysis techniques have been widely used in different credit evaluation processes. Therefore, classification problems are one of the applications that have gained serious attention over the past decades.

To slove the classification problem, there are many parametric discriminant methods proposed, the first is the linear discriminant function who is the oldest discriminant method initiated by Fisher in 1936 [14], it is the optimal combination which separates the averages from two groups.

This method of discrimination requires that the sample be distributed normally and that the variances-covariances matrixes of the two groups are homogeneous. The second method is the quadratic function suggested by Smith in 1947 [3], this method supposes the normality of the sample with heterogeneous variances-covariances matrixes. The last parametric method of discrimination is the logistic regression, which is an econometric method whose endogenous variable is binary, it requires neither the normality of the sample nor the homogeneity of the variances-covariances matrixes of the groups.

Recently, several linear programming models were proposed for resolving the classification problems by various authors such as Freed and Glover [10, 11, 12], Glen [8], Bajgier and Hill [16], Hasan and *al.,* [1], Markowski and Markowski [13], Gehrlein [17], Glover and *al.,* [6], Nath and Jones [15], Jones [4], Koehler and Erenguc [7] and Stam and Ragsdale [2].

Nevertheless, the all linear programming models suppose that the variables (or attributes) are measured with certainty. However, in an increasingly complex environment these variables can be imprecise, qualitative or linguistic. From where need for the recourse to fuzzy set theory (L. zadeh [9]). In this respect, we proposed a new approach, which consists in solving the classification problems via fuzzy linear programming, models based on ranking function proposed by F. Hosseinzadeh Lotfi and B. Mansouri [5].

The rest of the paper is organized as follows. In Section 2 we define the ranking function and these properties. In section 3, we present our new approach to solving the linear fuzzy classification problems. In section 4, the empirical study was carried out on a sample of 65 Tunisian firms, for which financial and account statements data are collected and 14 financial ratios are calculated. And in section 5, we give the concluding points and the future research.

2. Ranking Function

To deal quantitatively with imprecise data in classification

problems, the concept of fuzzy has been introduced. When variables are fuzzy, the objective function and the constraints of the decision model also become fuzzy.

In fact, we represent an arbitrary fuzzy by $\tilde{a} = (a^m, a^l, a^u)$ such that: a^m : medium value; a^l : lower value and a^u : upper value. According to F. Hosseinzadeh Lotfi and B. Mansouri [5], one of the most effective approaches to control all fuzzy numbers *F(R)* is to define a ranking function $\tau : F(R) \to R$ such as:

- $\tilde{a} \succeq \tilde{b}$ if and only if $\tau(a) \geq \tau(b)$
- $\tilde{a} \succ \tilde{b}$ if and only if $\tau(a) > \tau(b)$
- $\tilde{a} \approx \tilde{b}$ if and only if $\tau(a) = \tau(b)$

With $\tilde{a}$ and $\tilde{b}$ are fuzzy numbers.

We restrict our attention to linear ranking function which is a ranking function τ such that: $\tau(k\tilde{a} + \tilde{b}) = k\tau(\tilde{a}) + \tau(\tilde{b})$ for all $\tilde{a}$ and $\tilde{b}$ belonging to $\tau(R)$ and for $k \in R$. Indeed, for a fuzzy number $\tilde{a} = (a^m, a^l, a^u)$ we use the ranking function $\tau(\tilde{a}) = \frac{1}{2}(a^m + \frac{1}{2}(a^l + a^u))$.

Hence, for the triangular fuzzy numbers $\tilde{a} = (a^m, a^l, a^u)$ and $\tilde{b} = (b^m, b^l, b^u)$ we have:

$$(\tilde{a} \succeq \tilde{b}) \Leftrightarrow a^m + \frac{1}{2}(a^l + a^u) \geq b^m + \frac{1}{2}(b^l + b^u)$$

3. Proposed Methodology

Suppose there are n observations denoted by $X(i = 1, \ldots, n)$ each observation is characterized by p independent fuzzy variables denoted by $\tilde{x}_{ij}(j = 1, \ldots, p)$ for the ith observation. Suppose also that the observations are classified into two groups G_1 and G_2 containing, respectively, n_1 and n_2 observations such as: $G_1 \cup G_2 = G$ and $n_1 + n_2 = n$. The membership of observations in each group is known a priori. The objective is to find a rule that correctly classifies most imprecise observations. This rule enables us to find the group membership of any new imprecise observation. The classification rule is obtained from two stages. In stage 1 we determined a nonparametric function which reclassifies the observations; the second stage explains how to determine the membership of the observations which were not correctly classified at a stage. The objective is to minimize the total deviation of misclassified observations. These two stages are mathematically formulated as follows:

$$Min \sum_{i \in G_1} d_{1i}^+ + \sum_{i \in G_2} d_{2i}^- \quad (3.1)$$

Subject to

$$\sum_{j=1}^{p} (w_j^+ - w_j^-)\tilde{x}_{ij} + d_{1i}^+ - d_{1i}^- \approx b + 1, \ i \in G_1,$$

$$\sum_{j=1}^{p} (w_j^+ - w_j^-)\tilde{x}_{ij} + d_{2i}^+ - d_{2i}^- \approx b, \quad i \in G_2,$$

$$\sum_{j=1}^{p} (w_j^+ + w_j^-) = 1$$

$$d_{1i}^+, d_{1i}^-, d_{2i}^+, d_{2i}^- \geq 0, \quad i = 1, \ldots, n$$

$$w_j^+, w_j^- \geq 0, \quad j = 1, \ldots, p$$

In the above model w_j^+ and w_j^- represent, respectively, the positive and negative weights, d_{1i}^+ and d_{1i}^- represent, respectively, the positive and negative deviations of the observations of G_1. d_{2i}^+ and d_{2i}^- represent, respectively, the positive and negative deviations of the observations of G_2.

The model (1) is a fuzzy linear programming model, to obtain an equivalent deterministic model we use the ranking function τ. According to the property $\tilde{a} \approx \tilde{b}$ if and only if $\tau(a) = \tau(b)$ it is possible to change the previous model as follows:

$$Min \sum_{i \in G_1} d_{1i}^+ + \sum_{i \in G_2} d_{2i}^- \quad (3.2)$$

Subject to

$$\tau\left(\sum_{j=1}^{p} (w_j^+ - w_j^-)\tilde{x}_{ij} + d_{1i}^+ - d_{1i}^-\right) = \tau(b + 1), \ i \in G_1,$$

$$\tau\left(\sum_{j=1}^{p} (w_j^+ - w_j^-)\tilde{x}_{ij} + d_{2i}^+ - d_{2i}^-\right) = \tau(b), \quad i \in G_2,$$

$$\sum_{j=1}^{p} (w_j^+ + w_j^-) = 1$$

$$d_{1i}^+, d_{1i}^-, d_{2i}^+, d_{2i}^- \geq 0, \quad i = 1, \ldots, n$$

$$w_j^+, w_j^- \geq 0, \quad j = 1, \ldots, p$$

According to the properties of the ranking function (τ), we can replace the first two constraints by:

$$\sum_{j=1}^{p} (w_j^+ - w_j^-)\tau(\tilde{x}_{ij}) + d_{1i}^+ - d_{1i}^- = \tau(b + 1), \ i \in G_1,$$

$$\sum_{j=1}^{p} (w_j^+ - w_j^-)\tau(\tilde{x}_{ij}) + d_{2i}^+ - d_{2i}^- = \tau(b), \quad i \in G_2,$$

The final model in the first stage, when we apply the ranking function (τ), is formulated as follows:

$$Min \sum_{i\in G_1} d_{1i}^{+} + \sum_{i\in G_2} d_{2i}^{-} \tag{3.3}$$

Subject to

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] + d_{1i}^{+} - d_{1i}^{-} = b + 1, \quad i \in G_1,$$

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] + d_{2i}^{+} - d_{2i}^{-} = b, \quad i \in G_2,$$

$$\sum_{j=1}^{p}(w_j^{+} + w_j^{-}) = 1$$

$$d_{1i}^{+}, d_{1i}^{-}, d_{2i}^{+}, d_{2i}^{-} \geq 0, \quad i = 1,\ldots,n$$

$$w_j^{+}, w_j^{-} \geq 0, \quad j = 1,\ldots,p$$

Let $w_j^{*} = w_j^{+^{*}} - w_j^{-^{*}}\ (j = 1,\ldots,p)$ and b^{*} are the optimal solutions of the model above, the classification rule is as follows:

- $\frac{1}{2}\sum_{j=1}^{p}(w_j^{*+} - w_j^{*-})\left[x_{mj}^{m} + \frac{1}{2}(x_{mj}^{l} + x_{mj}^{u})\right] \geq b^{*} + 1, \ x_m \in G_1,$
- $\frac{1}{2}\sum_{j=1}^{p}(w_j^{*+} - w_j^{*-})\left[x_{mj}^{m} + \frac{1}{2}(x_{mj}^{l} + x_{mj}^{u})\right] \leq b^{*}, \quad x_m \in G_2,$

Otherwise x_m belongs to the area of overlap. In order to classify the observation x_m, the second stage begins. Before starting the second stages we define the following sets:

- $D_1 = \left\{ i \in G / \frac{1}{2}\sum_{j=1}^{p} w_j^{*}(x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})) \geq b^{*} + 1 \right\}$
- $D_2 = \left\{ i \in G / \frac{1}{2}\sum_{j=1}^{p} w_j^{*}(x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})) \leq b^{*} \right\}$
- $D_0 = \left\{ i \in G / b^{*} < \frac{1}{2}\sum_{j=1}^{p} w_j^{*}(x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})) < b^{*} + 1 \right\}$
- $C_1 = \{ i \in G / i \in G_1 \}$, $C_2 = \{ i \in G / i \in G_2 \}$, $G_1' = G_1 - C_1$, $G_2' = G_2 - C_2$

Hence, the model of the second stage is formulated as follows:

$$Min \sum_{i\in G_1'} d_{1i}^{+} + \sum_{i\in G_2'} d_{2i}^{-} \tag{3.4}$$

Subject to

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] \geq b + 1, \quad i \in C_1,$$

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] + d_{1i}^{+} - d_{1i}^{-} = c, \quad i \in G_1',$$

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] + d_{2i}^{+} - d_{2i}^{-} = c, \quad i \in G_2',$$

$$\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{ij}^{m} + \frac{1}{2}(x_{ij}^{l} + x_{ij}^{u})\right] \leq b, \quad i \in C_2,$$

$$\sum_{j=1}^{p}(w_j^{+} + w_j^{-}) = 1$$

$$b \leq c \leq b + 1$$

$$d_{1i}^{+}, d_{1i}^{-}, d_{2i}^{+}, d_{2i}^{-} \geq 0, \quad i = 1,\ldots,n$$

$$w_j^{+}, w_j^{-} \geq 0, \quad j = 1,\ldots,p$$

Let now $w_j^{*} = w_j^{+^{*}} - w_j^{-^{*}}\ (j = 1,\ldots,p)$ and c^{*} are the optimal solutions obtained in the second stage. Then the classification rule is as follows:

- $\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{mj}^{m} + \frac{1}{2}(x_{mj}^{l} + x_{mj}^{u})\right] \geq c^{*}, \ x_m \in G_1,$
- $\frac{1}{2}\sum_{j=1}^{p}(w_j^{+} - w_j^{-})\left[x_{mj}^{m} + \frac{1}{2}(x_{mj}^{l} + x_{mj}^{u})\right] \leq c^{*}, \ x_m \in G_2$

4. Empirical Study and Results

Our data base which was obtained from the "bourse des valeurs mobilières de tunisie (bvmt)" web site (http://www.bvmt.com.tn) based on a real data of 65 Tunisian firms divided into two groups. The first group (G_1) consists of 46 non-bankruptcy firms. The second group (G_2) constituted of 19 bankruptcy firms. Each firm is described by 14 financial ratios.

In this empirical study, we assumed that the 14 financial ratios characterizing the 65 firms are fuzzy triangular numbers.

The membership function of these variables are given in the Fig 1:

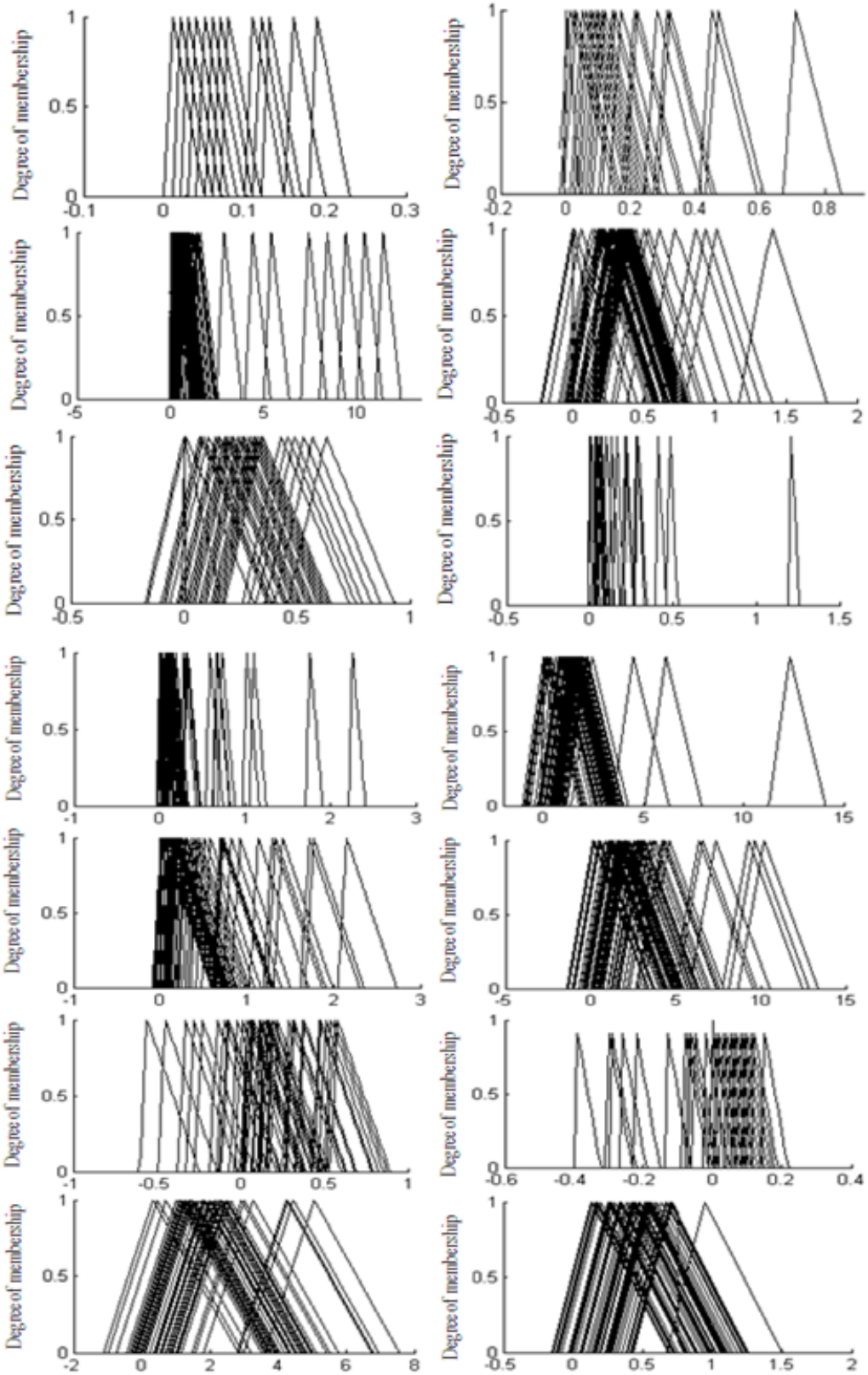

Figure 1. *The membership functions of the 14 financial ratios*

In the remainder of this section, we will exhibit the performance of the fuzzy classification linear programming using the ranking function defined in section 2.

The coefficients of the discriminant function and the value of the objective function of the first and the second stage of this model are given by the following table 1 (all results are given by the LINDO software):

Table 1. *The coefficients of the discriminant function of the first and the second stage*

	Stage1			Stage2			
w_1^+	0.0000	w_1^-	0.0000	w_1^+	0.0000	w_1^-	0.0000
w_2^+	0.0000	w_2^-	0.0000	w_2^+	0.0000	w_2^-	0.2432
w_3^+	0.0000	w_3^-	0.0000	w_3^+	0.0000	w_3^-	0.0001
w_4^+	0.0000	w_4^-	0.0000	w_4^+	0.0000	w_4^-	0.0895
w_5^+	0.0000	w_5^-	0.0000	w_5^+	0.0000	w_5^-	0.0000
w_6^+	0.0000	w_6^-	0.0000	w_6^+	0.0000	w_6^-	0.0000
w_7^+	0.0000	w_7^-	0.2267	w_7^+	0.0000	w_7^-	0.2742
w_8^+	0.0000	w_8^-	0.0000	w_8^+	0.0000	w_8^-	0.0053
w_9^+	0.0000	w_9^-	0.5627	w_9^+	0.0000	w_9^-	0.1211
w_{10}^+	0.0000	w_{10}^-	0.0765	w_{10}^+	0.0000	w_{10}^-	0.0332
w_{11}^+	0.0000	w_{11}^-	0.0000	w_{11}^+	0.0221	w_{11}^-	0.0000
w_{12}^+	0.0000	w_{12}^-	0.0000	w_{12}^+	0.0000	w_{12}^-	0.0000
w_{13}^+	0.1340	w_{13}^-	0.0000	w_{13}^+	0.0000	w_{13}^-	0.0000
w_{14}^+	0.0000	w_{14}^-	0.0000	w_{14}^+	0.0000	w_{14}^-	0.2113
b^*	-1.0884			b^*	-0.7412		
				c^*	-0.4118		
VOF	10.9998			VOF	0.1019		

VOF: Value of the Objective Function

The objective of the first stage is to identify the overlap between the observations based on the score given by the first discriminant function. Indeed, there is an overlap if and only if we have the classification score is between b and b^*+1 (i.e $-1.0884 < \sum_{j=1}^{14} w_j x_{ij} < -0.0884$).

The classification score showed the existence of an overlap between observations. The result of assigning observations at this stage showed that 19 observations belong to $G_1^{'}$ and 16 observations belong to $G_2^{'}$.

While, the objective of the second stage is to find a new discriminant function with a new threshold to reclassify misclassified observations.

Hence, the new classification rule is as follows:

If $\sum_{j=1}^{14} w_j x_{ij} \geq c^*(-0.4118)$ the observations belong to G_1

If $\sum_{j=1}^{14} w_j x_{ij} < c^*(-0.4118)$ the observations belong to G_2.

Moreover, it was noted that the value of the objective function in stage 2 has decreased compared to stage 1.

With regard to any classification problem, we must evaluate the performance of our results by referring to the criterion of the percentage of correctly classified. The classification result of our approach is given by the following table 2:

Table 2. *Correct classification rate of proposed method*

Group		The provided affectation Class		Total
		G1	G2	
Original Effective	G1	43	3	46
	G2	1	18	19
Rate	G1	93.478	6.522	100
	G2	5.264	94.736	100

According to table1,we can remark that only three non-bankruptcy firms thatarereportedasbankruptcyfirms (93.478% of the firmsin the first group are correctly classified) and one bankru ptcy firm siscl assified in the group of non-bankru ptcy firms (94.736% firms in the second group are correctl yclassi fied). Hence, the correct class if icationrate given by the proposed approachis 94.107%.There sultob tained by the proposed metho disver ysatis factory.

5. Conclusion

The aim of this paper is to evaluate a new approach for solving classification problems in the presence of fuzzy variables. In the first stage we have solved a first linear programming model to identify the overlap between the two groups. In the second stage, we solved a second linear programming model. While, the objective of the second stage is to find a new discriminant function with a new threshold to reclassify misclassified observations. To evaluate our approach, we calculated the rate of good classified obtained by the proposed method. This rate is equal to 94,107%.

The result is satisfactory and shows the ability of this procedure to solve classification some problems.

Given the relevance of this approach and its applicability to various classification problems, we think it would be interesting to show case our work:

- Adapting the developed method for other linear and nonlinear classification programming models;
- Extending the scope to other classification problems such as medical diagnostic, credit scoring etc.

References

[1] B. Hasan and al., "An experimental comparison of the new goal programming and the linear programming approaches in the two-group discriminant problems", Computers & Industrial Engineering vol 50, pp.296-311, 2006.

[2] C. T. Ragsdale and A. Stam, "Mathematical programming formulations for the discriminant problem: an old dog does new tricks", Decision Sciences vol 22, pp. 296-307, 1991.

[3] D. D. Smith and D. M. Whitt, "Estimating soil losses from field areas of claypan soils", Soil Science Society of America Proceedings Vol. 12, pp. 485-490, 1947.

[4] D. F. Jones and al., "A classification model based on goal programming with non-standard preference functions with application to the prediction of cinema-going behaviour", European Journal of Operational Research vol 177, pp.515-524, 2007.

[5] F. Housseinzadeh Lotfi and B. Mansouri, "The extended data envelopment analysis/discriminant analysis approach of fuzzy models", App Math Sci, Vol 2, N° 29-32, pp.1465-1477, 2008.

[6] F. Glover and al., "A new class of models for the discriminant problem", Decision Sciences vol 19, pp.269-280, 1988.

[7] G. J. Koehler and S. S. Erenguc S. S, "Minimizing misclassifications in linear discriminant analysis", Decision Sciences vol 21, pp. 63-85, 1990.

[8] J.J. Glen, "A comparison of standard and two-stage mathematical programming discriminant analysis methods", European Journal of Operational Research vol 171, pp.496-515, 2006.

[9] L. Zadeh, "Fuzzy sets", Information and Control Vol 8, pp. 338-353, 1965.

[10] N. Freed and F. Glover, "A linear programming approach to the discriminant problem", Decision Sciences vol 12, pp.68-74, 1981

[11] N. Freed and F. Glover, "Resolving certain difficulties and improving the classification power of LP discriminant analysis formulations", Decision Sciences vol 17, pp. 589-595, 1986.

[12] N. Freed and F. Glover, "Simple but powerful goal programming models for discriminant problems", European Journal of operational research vol 17, pp.44-60, 1981.

[13] P. Markowski and C. A. Markowski, "Some difficulties and improvements in applying linear programming formulations to the discriminant problem", Decision Sciences vol 16, pp.237-247, 1985.

[14] R. A. Fisher, "The use of multiple measurements in taxonomic problems", Ann. Eugenics, vol 7, pp.179-188, 1936.

[15] R. Nath and T.W. Jones, "A variable selection criterion in linear programming approaches to discriminant analysis", Decision Sciences vol 19, pp.554-563, 1988.

[16] S. Bajgier and A. Hill, "An experimental comparison of statistical and linear programming approaches to the discriminant problem", Decision Sciences vol 13, pp.604-618, 1982.

[17] W. V. Gehrlein, "General mathematical programming formulations for the statistical classification problem", Operations Research Letters vol 5, N°6, pp. 299-304, 1986.

[18] Alaleh Maskooki, " Improving the efficiency of a mixed integer linear programming based approach for multi-class classification problem", Computers & Industrial Engineering, Vol 66, Issue 2, pp. 383-388, 2013

[19] Leandro C. Coelho, Gilbert Laporte, " Classification, models and exact algorithms for multi-compartment delivery problems", European Journal of Operational Research, Vol 242, Issue 3, pp. 854-864, 2015

[20] Alireza Nazemi, Mehran Dehghan, " A neural network method for solving support vector classification problems", Neurocomputing, Vol 152, Issue 25, pp. 369-376, 2015

[21] Karim Ben Khediri, Lanouar Charfeddine, Slah Ben Youssef, " Islamic versus conventional banks in the GCC countries: A comparative study using classification techniques", Research in International Business and Finance, Vol 33, pp.75-98, 2015

[22] Xiao-bin Zhi, Jiu-lun Fan, Feng Zhao, "Fuzzy Linear Discriminant Analysis-guided maximum entropy fuzzy clustering algorithm", Pattern Recognition, Vol 46, Issue 6, pp. 1604-1615, 2013

[23] Xiaoning Song, Zi Liu, Xibei Yang, Jingyu Yang, "A fuzzy supervised learning method with dynamical parameter estimation for nonlinear discriminant analysis", Computers & Mathematics with Applications, Vol 66, Issue 10, pp. 1782-1794, 2013