Available online at http://ijdea.srbiau.ac.ir

Int. J. Data Envelopment Analysis (ISSN 2345-458X)

Vol.4, No.2, Year 2016 Article ID IJDEA-00426, 9 pages

Research Article

International Journal of Data Envelopment Analysis

Science and Research Branch (IAU)

Using DEA for Classification in Credit Scoring

Hoda Golshani[a*], Hadi Baghezadeh Valami [a], Alireza Davoodi[b]

(a) *Department of Mathematics, Yadegar - e- Imam Khomeini (RAH), shahr-e-rey Branch, Islamic Azad University, Tehran, Iran.*

(b) *Department of Mathematics, Neyshabur Branch, Islamic Azad University, Neyshabur, Iran*

Received 5 March 2016, Revised 29 June 2016, Accepted 15 July 2016

Abstract

Credit scoring is a kind of binary classification problem that contains important information for manager to make a decision in particularly in banking authorities. Obtained scores provide a practical credit decision for a loan officer to classify clients to reject or accept for payment loan. For this sake, in this paper a data envelopment analysis- discriminant analysis (DEA-DA) approach is used for reclassifying client to reject or accept class for case of real data sets of an Iranian bank branch. For this reason, two DEA models are solved. Also, the reject and accept frontiers and overlapping region among two frontiers are obtained. Then a goal programming problem is solved for finding co-efficients of the discriminant hyper-plane. The results are obtained from the samples are kept from the main dataset, clarify that the classified hyper-plane obtained from the used method provides an almost profitable classification for payment loan.

Keywords: Data Envelopment Analysis, Classification, Credit Scoring.

[*] Corresponding Author: golshani1400@gmail.com

1. Introduction

Data classification is a supervised learning problem which classes and class labels of each training data sets and costs of misclassification have been predetermined. The main goal in a classification problem is producing a separator which minimizes total costs for classifying data. For solving a classification problem two step algorithm is used. First step is recognition the overlapping region and second step is handling the overlapping. One of the most important features of this algorithm can be generalized and would provide simpler models for the separator. In other word, it is the process of organizing data into categories for its most effective and efficient use [1]. Application of classification is extended in various types in several literature such risk management Nanda et al 2001, pattern recognition Blue et al 1994, image processing Chellappa et al 1998 [2-4].

Credit scoring is a typical classification problem to categorize data in one of the predefined classes based on the number of classes related to that object. Most of the credit scoring problems are binary classification. The predefined label of classes are specially reject or accept in banking authorities. In other words, credit scoring models could provide important information to make managerial decisions. In such problems the scores are calculated based on ex-post information. Obtained scores provide a practical credit method to reclassify client for payment loans. [5-10].

For solving the classification problem there are several methods such as mathematical programming Duara Silva et al 1997; Freed 1986; Glen et al 1999 [11-13]. Among mathematical modeling, DEA is one of the most popular linear programming and frontier efficiency method which has advantages that can be suitable for solving the classification problem such: convexity for classes, linear piecewise classification frontier, the capability of developing the frontier when the information of a class is available, creation of an exclusive classification boundary for training data and the efficiency can be calculated with multiple inputs and multiple outputs. Some of these charastrictics are associated to the orientation and return to scales of DEA models [14]. In classification literature, these DEA based approaches are used for finding the discriminant hyper-plane is generally called DEA-DA methods. Sueyoshi 1999 has introduced DEA-DA as a non-parametric method for binary classification. Then, this seminal work and multi-criteria decision making (MCDM) were extended by Sueyoshi 2004, 2005, 2006 [15-23]. Also, Pendharkar et al 2011 have introduced a hybrid radial basis function using two BCC models in supervised learning phase for solving a binary classification problem. Their model was based on the feature of efficient and inefficient frontiers in DEA [24]. A basic issue in credit scoring system is to find more accurate methods for rejection or

acceptance clients for payment loans. Previous works have been limited on application of conventional DEA models in credit scoring. In addition, DEA-DA methods have been neglected to use for credit scoring models as supporting decision making in banking authorities. Our aim in this paper is according to the charastrictics of DEA-DA models come across two reject and accept frontiers and finding the overlapping region among two obtained frontiers. Moreover, for dealing with overlapped data set, the co-efficients of separator will be achieved by solving a goal programming on overlapping region. In addition, the costs of misclassification data are minimized, too. So, the overlapped data sets are reclassified more accurately. In particular, it seems that by using this approach for classification of clients, making a decision for payment loan is more profitable for bank. So, this study is focused on a special application of a DEA-DA algorithm for accepting or rejecting a client for payment or no payment loan in a national Iranian bank branch.

The rest of the paper is organized as follows. Section 2 some related researches by using DEA for credit scoring models will be reviewed. Section 3, using DEA-DA method in our case will be investigated. Section 4 numerical example is provided using real data set. Section 5 the paper is concluded.

2. Review of the Literature

In this section represent some documents addressing related issues. One of the preliminary attempt for a credit application of DEA have been conducted by Troutt et al 1996 [25]. Pendharkar 2002 had used DEA for addressing the inverse classification problem for predicting bankruptcy of firms [26]. Furthermore, DEA were mentioned by Samreen 2012, Their model could be handled negative data. Hanafizadeh et al 2014 [7] have used DEA for measuring the efficiency of mutual funds. In their model every mutual fund has considered as a DMU. Toloo et al 2015 have presented a hybrid approach for the largest private bank in Iran. Their proposed model handled negative data [27]. Abdou 2007 [4], have proposed new loan programs basedon DEA. A seven step methodology base on DEA have proposed by Emel et al 2003 [6] for credit scoring for the commercial banking sector.

3. Implementation of Dimensionality Reduction in Credit scoring Modelling

In this section, we are focused to investigate the application of a DEA-DA approach that proposed by Pendharkar et al 2012 [28] for modelling a credit scoring problem of real data set. In using method, finding the overlapping region will be considered as dimensionality reduction. So, first let us define dimensionality reduction problem as:

Definition1. The problem of dimensionality reduction includes the compression of data matrix D whose dimensions $n \times m$ (n>2 and m>2) to matrix C with less dimensions, that matrix C contains the most important

information of the main matrix [1]. We remind that for high-dimensional datasets (i.e. with the number of dimensions more than 10), need to apply algorithms for dimensionality reduction. For this reason, first we need to extract the features of each class of data sets. In this section, by solving two DEA models for accept or reject class, a two step algorithm will be used for implementation real credit scoring in practice. From the perspective of orientation, using DEA models are output and input oriented, for recognition accept and rejct class, respectively. Each client can be considered as DMU due to belong to reject or accept class. Model (1) deal with each selected feature as output and the model is output oriented. The frontier has achieved from solving Model (1) is called acceptance frontier. Model (2) handling each selected feature as input and the envelopment form is input oriented. The procedure of finding the overlapping region is a two step algorithm. From the expert point of views, first the real data set *P* is partitioned into two sub-sets P^A and P^R where $|P^A| = J$ and $|P^R| = K$ consist acceptance and rejection class, respectively. For DMUs which are in P^A , the linear programming Model (1) is solved $|P^A|$ times. DMUs which have $\varphi^* = 1$, are in the acceptance reference set of p^A where X_{mj} in Model (1) is *m*th feature related to *j*th DMU [28]. So, by solving Model (1) efficient frontier that associated the feature of DMUs underlying on the frontier, are completely efficient. This frontier is called accept frontier.

$$
\begin{aligned}
&Max\ \varphi^t \\
&s.t. \\
&\sum_{j=1}^{J} \mu_j x_{mj} - \varphi^t x_{mj} \geq 0\ , m = 1, \dots, M, j = 1, \dots, J \\
&\sum_{j=1}^{J} \mu_j = 1 \\
&\mu_j \geq 0 \quad \forall j\ \forall m
\end{aligned} \tag{1}
$$

$$
\begin{aligned}
&Min\theta \\
&s.t. \\
&\sum_{k} \lambda_k x_{mk} - \theta x_{mk} \leq 0, m = 1, \dots, M, k = 1, \dots, K \\
&\sum_{k} \lambda_k = 1 \\
&\lambda_k \geq 0 \quad \forall k
\end{aligned} \tag{2}
$$

The Model (2) is solved at $|\mathrm{P^R}|$ times for DMUs which are in $\mathrm{P^R}$set. The reference set of R^{R} includes clients which have $\theta^* = 1$. In Model (2), $\mathrm{x_{mk}}$ is the *m*th feature related to the DMU_{k} The obtained inefficient frontier by solving Model (2) is called reject frontier. The set of $P^{SV} = R^A \cup R^R$ and P^{SV} is the reference set of reject and accept classs. This set contains the main feature of both classes. In addition, $|P^{SV}|$ determines a lower bound for reducing the dimension of real data set. Because of P^{SV} is the smallest set that consists the main feature of each class. Between two frontiers the overlapping region is situated. For seeking the overlapping region $\forall x_t \in \{P - P^A \cup R^A\}$, Model (1) is solved $|P - P^A| + |R^A| = D$, times. If the cases of reject class change R^A then eliminate from the set of x_ts otherwise they will remain in the set. The

reminder cases are misclassified data from the rejection sets that are overlapped to the accept class. On the other hand, the misclassification data of accept class that have been classified in reject class obtain by solving Model (2) $|P - P^R| + |R^R| = U$, times. If the reject case changes R^R , it must eliminate. Otherwise it is added to overlap data set. The notation of the overlapped data set is S^V.

Let us extend the implementation of the method to real application in credit scoring. In our case study, the initial sub-sets of reject and accept classes from the main data set have been made by a validation system of national public banks. Since, loan manager need to make a decision in order to repay or not repay loan are based on scoring of various items achieve from clients' information. Due to the expert point of view, the score of intangible and tangible assets of real client is considered as the feature x_1 and the history of performing loans to the bank considered as the feature x_2. These two main features are used for our real computations for the realization of DEA-DA method for designing a supporting decision making for national validation system.

Under the monotonicity assumption, If $S^V - \{R^A \cup R^R\} \neq \varnothing$ then the following propositions are held in the overlapping region, for proofs see [28].

***Proposition* 1.** Classification area which is obtained by linking element of P^{SV} is convex.

***Proposition* 2.** All elements of S^V have been located inside convex region have achieved by linking elements of R^A and R^R.

Proposition 3: The number of misclassification for each linear discriminant function located in the generated convex region obtained from linking elements of R^A and R^R is equal to at least one.

The main goal in the second step is encountering the overlapping region. For credit decision making considering the cost of misclassification for classifying is meaningful. For this sake, a goal programming model is used considering misclassification costs in the objective function. This model is solved on the dimensionality reduction region of training data. Corresponding to the proposition (1) the overlapped region is convex and according to the proposition (2), all elements of S^v are inside this region. The goal programming is as:

$$\text{Min}\left(c(^2/_1) \times \sum_{i \epsilon s_1} d_i^+\right) + \left(c(^1/_2) \times \sum_{i \epsilon s_1} d_i^-\right)$$

s.t 0

$$\sum_{j=1}^{m} \alpha_j X_{ij} + d_i^+ - d_i^- - \alpha_0 = 0 \quad , \quad \forall i \epsilon s_1$$

$$\sum_{j=1}^{m} \alpha_j X_{ij} + d_i^+ - d_i^- - \alpha_0 = -\varepsilon \ , \quad \forall i \epsilon s_2$$

$$d_i^+ \geq 0 \ , d_i^- \geq 0 \ , \forall i \epsilon \{1, \dots, u\} \qquad (3)$$

$$infinitesimally \quad non - archemedian$$

$$\alpha_j \geq \varepsilon \ , j \ \epsilon \{1, \dots, m\} \ , \varepsilon$$

$$\alpha_0 \qquad unrestricted$$

Where $c(^1/_2) = 2$ and $c(^2/_1) = 1$ are misclassification costs in the group S_1, S_2 and the variable d_i^+, d_i^- are positive and negative deviations from the linear discriminant hyper-plane, respectively. α_j s are the co-efficients

of discriminant function corresponding to each feature X_{ij} . S_1, S_2 are accept and reject data set, respectively. Due to the obtained hyper-plane the misclassification data are reclassified as:

$$A(X) = \sum_{j=1}^{m} \alpha_j X_i \, , \frac{\varepsilon A}{\varepsilon x_j} \geq \varepsilon > 0 \, , \forall \, j \, \epsilon \{1, \dots, m\}$$

and coefficients of β_j are always positive. While X_{ij} must always be positive, Model (3) is solved for data inside the overlapping region.

If $d_i^{-^*} > 0$ were classified inside S_1 group and if $d_i^{+^*} > 0$were classified inside S_2 class. The classes of testing and new data are determined in such manner.If $-\alpha_0 + \sum_{j=1}^{m} \alpha_i^* X_{ij} \geq 0$ then the data will fall into S_1group otherwise they will be classified into S_2 group. For the sake of supporting decision making in banking activites two groups are accept or reject for payment loan. So, profitability for bank is considered as main goal in empirical part of our study.

4. The application in credit scoring

In this section, the application of DEA-DA model in credit scoring is clarified.For the sake of comparison with national validation system scores for reclassifying clients a DEA-DA method will be used. In national validation system demands of real clients with scores more than fifty are accepted. As mentioned in foregoing part for measuring the scores feature x_1 is the score of intangible and tangible assets of real client and how to performing loan is considered as the feature x_2. The class of zero is associated with rejection and class of one indicates acceptance of loan demanding. The scores corresponding each feature and the initial class of 24 clients based on the validation system scores is illustrated in Table 1.

Via the class information of clients in Table1 associated with validation system scores $P^A = \{1, \dots, 14\}$ and $P^R = \{15, \dots, 24\}$. The results of dimensionality reduction for the above real example obtain as follows:

$$S^T = \begin{pmatrix} 13/68 & 0/38 & 1/09 & 13/86 & 0/34 & 0/47 & 7/09 \\ 0/830 & 7/09 & 7/09 & 5/88 & 9/40 & 7/09 & 5/88 \end{pmatrix}$$

$$P^T = \begin{pmatrix} 0/34 & 0/47 & 13/68 & 0/38 & 1/09 & 13/86 \\ 9/40 & 7/09 & 0/83 & 7/09 & 7/09 & 5/88 \end{pmatrix}$$

Table1.The information of 24 real clients of an Iranian bank

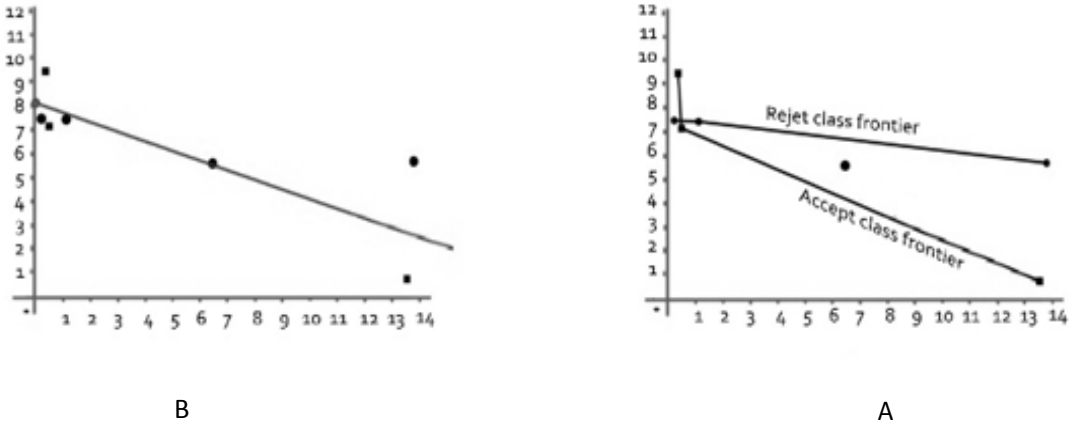
Number of Client	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
X_1	13/68	10/95	13/68	2/72	2/74	0/34	13/68	6/84	0/47	13/68	5/47	13/68
X_2	11/82	11/82	11/82	7/09	11/62	9/4	7/09	7/09	7/09	7/09	7/09	7/09
Class	1	1	1	1	1	1	1	1	1	1	1	1
Number of Client	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
X_1	13/68	0/74	1/79	1/79	6/45	10/79	0/38	2/74	0/38	1/09	0/0	13/68
X_2	0/83	8/33	8/33	5/88	5/88	5/88	1/06	7/09	5/88	7/09	5/88	5/88
Class	1	1	0	0	0	0	0	0	0	0	0	0

Where sets S^V={6, 9, 12, 17, 21, 22, 24}, P^v={6, 9, 12, 19, 22, 24} are overlapping clients and the reference set of reject and accept class, respectively C^V and S^v and two rejection and acceptance frontiers are demonstrated in Figure (1-A). As it is evident, S^V- $P^{sv} \neq 0$ holds, practically. The co-efficients of the discriminant hyper-plane are obtained by solving a goal programming on dimensionality reduction region. The hyper-plane is obtained as $D(X)=-0/147+0/01x_1 +0/123x_2$. By using this discriminant hyper-plane misclassified data are reclassified. Figure (1-B) illustrate the discriminant function for reclassifying misclassified clients. In practice, it is evident that rejection the demand of 9th and 12th client will be more profitable for the bank. The cost of classifying of those incorrectly samples which have been held in primary data matrix D is about $2/63\times10^{-12}$ that achieve by Model (3).

The minimum cost obtained for the held data of above main data set reveals that the line of discriminant function is capable to highly extend to test samples while classify training samples more accurately and this is the same thing which expect from a decision support systems, specially in credit problem because to keep the profitability in banking to make a decision.

5. Conclusion

The main contribution of this paper was special implementation of DEA-DA in credit scoring. For the sake of comparison to the national validation system for payment loan a decision supporting system was suggested to make more precisely decisions. For reclassifying at the empirical part, on reducing dimension a goal programming was used. The results of data samples held from the main data set indicate that the classification hyper-plane obtained by this method is a perfect separator.



B A

Figure(1).An illustrate accept , reject frontier and overlapping region B. The co-efficient of discriminant hyper-plane are obtained by solvig goal programming

The future study framework can be focused on applying this algorithm for big data set. Because of dimensionality reduction prevents from computational complexity in banking authorities. So, the domain of study can be developed for the other Iranian banks that use the same validation system for payment loan.

Acknowledgements

The authors would like to thank the anonymous reviewers and the editor for their insightful comments and suggestions.

References

[1] Duda. R. O. Hart. P. E and Strok. D. H. Pattern Classification (2nded) New York: Wiley Inter Science; (2001).

[2] Blue. J. L. Candela GT, Grother. P. J, Chellapp (1994). Evaluation of Pattern Classifier for finger print and OCR applications, Pattern Recognition. 27(4): 485-501.

[3] Nanda S, Pendharkar, P. C (2001). Development and comparison of analytical techniques for Predicting in solvency risk. System in accounting finance and Management, 10:155–168.

[4] Chellappa. R, Fukushima. K, Katsagelos, A. K. Kung S.Y. Lecun, Y, Nasr abadi N.M, et al (1998). Special issue on, Applications of Artifical Neural Network to Image Processing. IEEE. Transactions on Image Processing, 7 (8): 59-71.

[5] Abdou,H, El-Masry, A, Pointon, J (2007). On the applicabillity of credit scoring models in Egyptian banks, Banks and Bank Systems, 2(1): 4-18.

[6] Emel, A.B, Oral. M (2003). A credit scoring approach for the commercial banking sector, Socio-economic planning science, 37(2): 103-123.

[7] Hanafizadeh, P, Khedmatgozar, H.R Emrouznejad, A Derakhshan M (2014). Neural network DEA for measuring the efficiency of mutual funds, International journal of applied decision sciences, Volume 7(3): 255 - 269.

[8] H. Min J, Chan Lee Y (2008). A practical approach to credit scoring. Expert Systems with Applications 35(4):1762–1770.

[9] Thomas L. C, Edelman D.B. Crook J.N, Credit Scoring and Its Applications Philadelphia: Siam, (2002).

[10] Akgöbek Ö, Yakut E (2014). Efficiency measurement in Turkish manufacturing sector using data Envelopment Analysis (DEA) and artificial neural Networks (ANN), Journal of Finance and Banking. 1(2): 36-47.

[11] Duara Silva A.P. Stam (1997). A mixed-integer programming algorithm for minimizing the training sample mis classification cast in two – group classification Annuals of Oprations Research, 129-157.

[12] Freed. N. Glover . F (1986). Evaluating alternative Linear programming models to solve the two group discriminant problem. Desicion Science, 17: 151- 162.

[13] Glen J.J. Integer programming method for

normalization and variable selection in mathematical programming discriminant analysis models (1999). Journal of the operational Research Society, 50: 1043-1053.

[14] Banker R.D, Charnes A, Cooper W.W. (1984) Some models for estimating technical and scale inefficiency in data envelopment analysis. Manag Sci 30:1078–1092.

[15] Sueyoshi T (1999). DEA-discrimininant analysis in the view of goal programming: European journal of operational Research. 115: 564–582.

[16] Sueyoshi T, Hwang . S.N, A use of non parametric test of DEA – DA A methodological comparision, Asia Pasific journal of operational Research, (2004), 21 (forthcoming).

[17] Sueyoshi, T. Mixed integer programming approach OF Extended DEA- discriminant analysis. European journal of operational Research, (2004)152, 45–55.

[18] Sueyoshi T. Sekitani K. Return to scale in dynamic DEA. European journal of operational Research, (2005), 161:536–44.

[19] Sueyoshi T. DEA- discriminant analysis: methodological comparison among eight discriminant ananlysis approaches. European journal of operational Research, (2006), 169, 247-72.

[20] Pendharkar, P. C. Nanda. S, A misclassification Cost minimizing evolutionary neural classification approach Naval Research Logistics, (2006), 53 (5), 432–447.

[21] Nakayama. H. Ka Gaku. N. (1998). Pattern classification by linear goal programming and its extensions. Journal of Global optimization, 111-126.

[22] Pendharkar. P.C. Trout. M.D (2012). DEA based preprocessing for minimum decisional efficiency linear case valuation models. Expert systems with applications. 39: 435–9442.

[23] Troutt, MD (1995). A Maximum decisional efficiency estimation principle, Management science, 41(1): 76-82.

[24] Pendharkar. P. C. A (2011). Hybrid radial basis function and data envelopment analysis neural networks for classification. European journal of operational research, 38 (1): 256–266.

[25] Troutt, MD, Rai. A, Zhang A (1996). The potential use of DEA for credit applicant acceptance system. Computer and operations Research, 23(4): 405-408.

[26] Pendharkar. P. C. (2002). A Potential use of DEA for inverse classification problem, omega, 30(3): 243–248.

[27] Toloo M, Zandi. A, Emrouznejad. A (2015). Evaluation efficiency of large-scale data set with negative data: an artificial neural network approach, Supercomput 71: 2397–241.

[28] Penharkar , P, C, Troutt M. D (2011). DEA based dimensionality reduction for classification problem satisfying strict non satiety assumption. European Journal of operational Research, 212(1): 155-163.