Optimization
Iranian Journal of Optimization
Volume 8, Issue 2, 2016, 59-69
Research Paper
Online version is available on: www.ijo.iaurasht.ac.ir
Islamic Azad University
Rasht Branch
E-ISSN:2008-5427

An Integrated DEA and Data Mining Approach for Performance Assessment

Alireza Alinezhad*

Associate Professor, Faculty of Industrial and Mechanical Engineering, Qazvin Branch, Islamic Azad University, Qazvin, Iran

Received: 21 September 2016
Accepted: 03 November 2016

Abstract

This paper presents a data envelopment analysis (DEA) model combined with Bootstrapping to assess performance of one of the Data mining Algorithms. We applied a two-step process for performance productivity analysis of insurance branches within a case study. First, using a DEA model, the study analyzes the productivity of eighteen decision-making units (DMUs). Using a Malmquist in-dex, DEA determines the productivity scores but cannot give details of factors depend on regress and progress productivity. The proposed model presents anew latent variable radial input-oriented technology and simultaneously reduces inputs and undesirable outputs in a single multiple objective linear programming. On the other hand, classification and regression tree allow DMU to extract rules for ex-ploring and discovering meaningful and hidden information from the vast data-bases. The results provide a set of rules that can be used by policy makers to explore reasons behind the progress and regress productivities of DMUs.

Keywords:
Data envelopment analysis
Classification and regression tree
Bootstrapping
productivty
Malmquist index

***Correspondence E-mail**: alalinezhad@gmail.com

INTRODUCTION

Evaluation of efficiency and productivity of decision-making units (DMUS) such as banks, insurance companies, universities, and so on, using multiple inputs and outputs, is wrapped. There are some researches showing the importance of process in assessing the performance of a firm(Charnes et al., 1978). In recent decades, data envelopment analysis (DEA) has gained significant growth as a powerful managerial tool for efficiency measurement and it has been used wide spread for evaluating the efficiency and productivity of public and private sectors. DEA applies inputs/outputs variables to create an efficient boundary from a series of considered DMUs.The efficiency of each DMU is computed by measurin the distance of the DMU from the efficient frontier.Similarly, Malmquist productivity index is used to evaluate technology change and its effect on inputs and outputs. It is defined as the maximum factor by which inputs in one period could be reduced to produce the same output in a second period (Pille & Paradi, 1997).

Data mining techniques extracting patterns from large databases have recently become prevalent. Data mining is a method usually used to find out meaningful communication and rules by systematically breaking down and subdividing the information in the data (Chen et al., 2003). Breiman et al. (1984) introduced a hierarchical sequence of decision nodes approach called classification and regression tree (CART) algorithm in which each node in a tree strikes one decision at a time until a final node is achieved. Various variables are utilized and a special variable enters the computation only when it is required at a special decision node, and only one variable is utilized at each decision node (Clark & Pregibon, 1992).

In this paper, a two-stage performance evaluation applying DEA, anon-parametric method for efficiency evaluation, and a CART tree, a non-parametric data mining method for classification and regression, is presented to evaluate the performance of insurance branches of Iran insurance cooperation. Productivity scores provide valuable data for the performance assessment of insurance branches while the CART tree determines further facts that have not been recognized in prior studies. Sohn and Moon (2004) proposed an approach that can be effectively used for foreshadowing the scale of new technology commercialization projects using the Decision Tree (DT) of DEA results. Lee and Park (2005) applied a combination of these methods for classified profitable customer.Seol et al. (2007) proposed an approach that enables firm's manager to find inefficient service units in a firm-level and inefficient process in a service unit-level by using an integrated form of DEA and DT.Using a combination of DT and CART DEA models, Emrouznejad and Anouze (2010) evaluated the performance of Arabic banks and finally, Samoilenkoand Osei-Bryson (2013) demonstrated a DEA-centric decision support system (DSS) in order to propose how to assess and manage the relative performances.The article organized as follows: Section 2 gives summery debate of DEA and the CART tree. This is followed by our proposed methodology of a DEA/CART approach in Section3. Section4presents and explains an empirical analysis of suggested framework DEA/CART. Finally, Section 5 presents the conclusions and points for future works.

RELATED WORK

In this section, we review the literature in three main sectors consist of DEA, CART, and hybridized DEA and data mining techniques.

Brief on DEA

DEA is a non-parametric technique for assessing the efficiency of DMUs with multiple inputs and outputs proportions of weighted outputs to weighted inputs, and to define the relative efficiency comparing with other DMUs. DEA was initiated by Charnes et al. (1978) to demonstrate how to change a fractional linear measure of efficiency into a linear programming model. The Slacks Based Model (SBM) is Pareto-Koopmans in which there is no test for choosing the best specification or model in DEA as noted by Bretholt and Pan (2013). The proposed methodology in this study called the LatentVariable Model is anew Latent Variable Model (LVM) radial input-oriented technology that is closely associated with the Koopman Efficient Slacks Based Model. The latent variable technology si-

multaneously reduces inputs and undesirable outputs in a single multiple objective linear programming.

Technical efficiency in the production processes such as insurance corporations turns inputs into outputs (administrative costs, insurance costs and number of agents as inputs and the value of loan payments and the income from insurance premium as desirable outputs and compensation value as undesirable output). The relationship between inputs and outputs can be shown by a production function which demonstrates the maximum outputs possible for a given level of inputs.

The proposed model, non-parametric method of undesirable outputs with weak disposable inputs technology Since the technology included the undesirable outputs, Bretholt and Pan (2013) introduced a method that could be built on the following principles:

Using inputs (x_{pj}, p =1, 2,…, P) and producing Q desirable outputs (y_{qj}, q =1, 2,…, Q) and R undesirable outputs (z_{rj}, r =1, 2,…, R).

Assume that there are J branches of an insurance corporation.

Latent Variable technology uses a Radial Input Model in association with weak disposability applied to aggregate inputs. The weak disposable inputs aggregate inputs, x_{pj} are reduced by the direct input reduction objective, α as follows:

$$\frac{\sum_{j=1}^{J} z_j x_{pj}}{\alpha_0 x_0} = 1 \qquad (1)$$

All axioms, along with a shrinkage factor for undesirable factors are essential for the formation of LV model [1] . Hailu and Veeman claimed that if the specified undesirable inputs or outputs are replaced with the $X = \sum_{j=1}^{J} z_j x_j$ Equation, the resulting DEA model shows weak disposable inputs. (Fare et al., 2008).

Let us correctly specify the latent variable model now. To generalize the latent variable input model suppose two time periods exist, K and L, and that vectors with P inputs, Q outputs, R undesirable outputs, and J DMUs are given as shown Eq. 2.

$$\begin{aligned}
&VRS\,LV\,Min\alpha:\\
&\{\forall DMU \,||\, j=1,2,...,J : t=K,L\}\\
&s.t.\sum_{j=1}^{J} z_j x_{pj}^{t} = \alpha x_{p0}^{t}, \quad p=1,2,...,P\\
&\sum_{j=1}^{J} z_j y_{qj}^{t} \geq y_{q0}^{t}, \quad q=1,2,...,Q\\
&\sum_{j=1}^{J} z_j u_{rj}^{t} \leq \lambda u_{r0}^{t}, \quad r=1,2,..,R\\
&\sum_{j=1}^{J} z_j = 1\\
&z_j \geq 0\\
&Latent\,Vaiable \quad 0 \leq \lambda = \frac{\sum_{j=1}^{J} z_j u_{rj}^{t}}{u_{p0}^{t}} \leq 1\\
&\{\forall DMU \,||\, r=1,2,...,R : p=1,2,...,P\}
\end{aligned} \qquad (2)$$

In this article, a new model for evaluating the efficiency of outputs as inputs is proposed. Considering a shrinkage factor for the undesirable outputs, the dispersion between periods is studied using the Malmquist Productivity Index (MPI). After determination of the DMUs efficiencies, their productivities will be specified for the periods of 2008-2009 and 2009-2010 based on the following formulae.

The MPI is used to assess technology changes and change effect on the inputs and outputs(Zhu, 2004; Malmquist,1953).

The largest factor MPI is defined by the inputs that can be reduced in one period and determine the same output's production in a second period.

Suppose the production technology in the K period when the main reduction coefficient is as follows and the target values are in the L period.

$$\lambda_j^k(X_j^L, Y_j^L, U_j^L) \qquad (3)$$

In view of the reference technology in the K period, by changing to $\lambda_j^k(u_j^L)$, the efficiency of undesirable output in the L period can be calculated.

Similarly, $\alpha_j^K(X_j^l)$ is the efficiency of inputs in the L period when the reference technology is calculated in the K period.

In general, the MPI can be decomposed into two components of technical efficiency changes and (efficiency) production frontier shifts.

The main symbol is combined with the MPI components in Eq. 2. The overall objective in this decomposition method is reducing α and the hid-

[1] *Latent Variable Model*

den variable of λ.

The partial correlation coefficients are resulted from a dense LVM when it helps to minimize the variance in the model as much as possible.

$$\frac{U_j^L}{U_j^K}=\left(\frac{(U_j^L/(\lambda_j^K(U_j^K)\lambda_j^L(U_j^K))^{1/2})(1/X_j^L)}{(U_j^K/(\lambda_j^K(U_j^L)\lambda_j^L(U_j^L))^{1/2})(1/X_j^K)}\right)\times$$
$$\left(\frac{(X_j^L/(\alpha_j^K(X_j^K)\alpha_j^L(X_j^K))^{1/2})(1/Y_j^L)}{(X_j^K/(\alpha_j^K(X_j^L)\alpha_j^L(X_j^L))^{1/2})(1/Y_j^K)}\right)\times$$
$$\left(\frac{\lambda_j^K(U_j^K)}{(\lambda_j^L(U_j^L)}\right) \quad (4)$$

In general, the introduction of the DEA hidden variable technology is a first step towards the analysis of undesirable outputs and the consideration of external effects on the company and the society. Using the dense hidden variable reduction model, this model theoretically presents the production of simultaneous reduction of undesirable outputs and inputs through causal relationships. Eq. 4 shows DMUs productivity results; accordingly, MPI>1 indicates progress,MPI=1 shows no change, and MPI<1 is indicative of regress during the period.

Brief on CART

The data mining technique allows DMUs to discover significant information that had previously been hidden in large databases.

CART, a decision making tree normally used in data mining processes, has been developed in 1984 by Breiman and improved in 1996 by Ripely. The problem is illustrated by a decision making tree so that each non-leaf node is associated with one of the decision making variables, each branch of a non-leaf node is associated with a subset of the decision making variables values, and each leaf node is linked to a target variable (the dependent variable) value. Each leaf is associated with a target variable's mean value; therefore, this tree can be an alternative to continuous linear models for solving the problems of regression and logistic regression analyses of classified data (Clark et al., 1992).

In general, CART trees have some advantages over the regression models. First of all, a model created by a tree is more plausible and relatively simpler for non-statistic a interpretations (Breiman et al., 1984; Han et al., 2001). Secondly, its non-parametric nature indicates that it has been made by the independent variables values without any pre-assumption. Therefore, CART trees can handle numerical data that have high Skewness or are multifaceted as well as categorical predictors with sequential and non-sequential structures. Thirdly, compared with the regression models, CART trees have more sophisticated methods to deal with missing variables. In regression, data that contain any missing value will be automatically deleted. Hence, CART trees can be created even when some independent variables are not recognizable for a number of DMUs. Finally, CART trees are somewhat an automatic machine learning method. CART trees present computational efficiency in order to need less time for computation and storage of the data.

In creation of a CART tree, data set is usually divided into two parts: the training data set and the test data set (Hann et al., 2001; Hand et al., 2001).

Then they undergo two main processing phases of growth and pruning.

In the development stage, a CART tree is constructed from a set of training data. In this phase, each leaf node is associated with a class.

In order to avoid over-fitting, the produced CART tree is improved in the pruning stage. At this point, the CART tree is evaluated for being a sub-tree with the lowest error rate for the set of experimental data. Ripley et al. (1996) and Hand et al. (2001) have represented an exact algorithm for CART trees. In these articles, CART trees analyses are presented to explore and evaluate both internal and external factors (e.g. the number of claims paid, the number and qualifications of staff, level of the branch, number of loans, number of insurance policies, etc.) that are all influential in the efficiency of insurance branches.

Efficiency and productivity scores derived from the DEA constitute target values of a tree. Thus, DMUs are divided into two categories of progressive productivity and regressive productivity.

Combination of DEA with data mining technique

The former DEA researchers mainly focused on functional evaluations and regulations; therefore, only a few cases of combining DEA with technique of data mining have been reported. For example, Sohn and Moon (2004) examined the possibility of combining decision trees with DEA in Research and Development (R&D) projects (i.e. when a company is trying to discover new knowledge or develop new technology) (Barr et al., 1994).DEA researchers have focused mostly on the assessment and control of past performance and only few attempts have been reported to combine DEA with data mining such as prediction of bank failure prediction and failure of Credit Union (Pille et al., 1997). However, no study has been done concerning the combination of DEA and CART tree in evaluating the efficiency and productivity of insurance branches.

MATERIALS AND METHODS

Combining DEA with CART tree

The proposed CART tree in this study includes four main components:

The first component, is the output (dependent) variable. Based on the independent (predictive) variables, this variable is used to predict.

In this study, the output variable is the obtained productivity scores that have been divided into three groups of progressive productivity (target > 1), regressive productivity (target < 1), and without change productivity (target= 1).

The second component is the independent (predictive) variables. The number of independent variables is related to the purpose of investigation.In this case, the independent variables are external and internal factors (Table 1).The third component is the set of training data, which includes both output and independent variables values coming from a group of DMUs we want to predict.

The fourth component is the test or the set of additional data coming from specific DMUs that require more precise prediction. This test data set may not exist in practice. It is normally believed that a test data set is required to enforce the decision laws; however, it is not always necessary to determine the efficiency of the decision laws. Using DEA/CART, the evaluation process of efficiency and productivity of insurance branches is presented in Fig.1. As shown in figure, first DEA is applied to measure the efficiency and productivity of each branch with three inputs (administrative costs, insurance costs, and the number of branches) and three outputs (revenue from insurance premiums, the loan payments, the compensation payments). According to these results, the branches will be divided into three groups of efficient, inefficient, and without change branches.

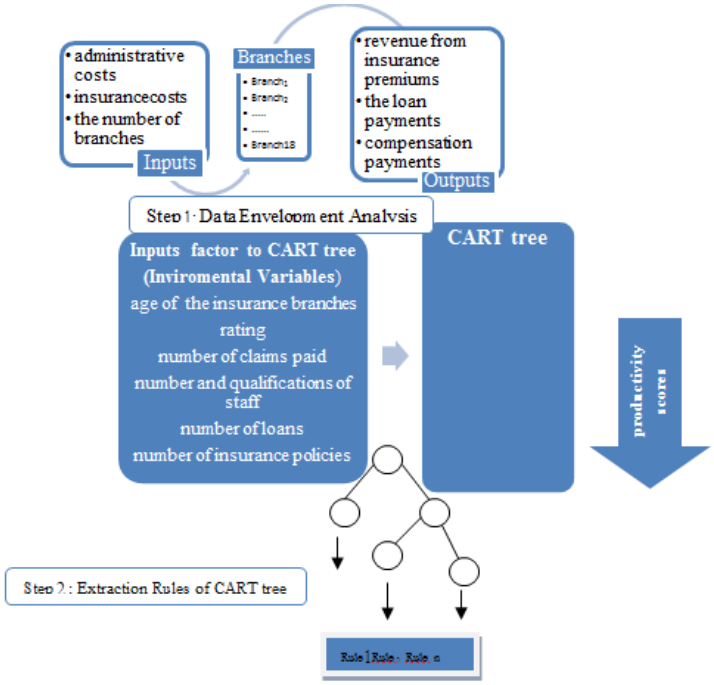

Fig.1. DEA/CART methodology for assessing Iran Insurance

Table 1: Input factors in the CART tree

Variable	Variable type	Min	Max	Mean	Std
Age of the insurance Branches	Numerical	1386	1324	1372	12.936
Level of the branch[a]	Categorical	1	3		
Geographical branch[b]	Categorical	1	5		
No. of staff	Numerical	10	159	48.66	39.065
Qualifications of staff[c]	Categorical	1	5		
No. of loan	Numerical	2	1007	354.963	325.191
No. of insurance policies	Numerical	12	6894	1251.56	1780.511
No. of claim paid	Numerical	6	779	263.463	247.7242

a_1,Assembled;$_2$,Super;$_3$,Level $_1$
b_1,North;$_2$,South;$_2$, East;$_4$, West;$_5$, Center
c_1,Diploma;$_2$, Advance Diploma;$_3$, BA;$_4$, MA;$_5$, Phd

In the next stage, insurance-related environmental factors such as age of the insurance branches, their ratings, and the number of issued insurance policies are considered as inputs to the CART tree analysis while productivity scores, obtained in the first phase, are regarded as the outputs (Table 1).Clearly this is a general framework applicable in conducting all types of analyses in every organization including insurance companies and banks.If this method is used for other purposes, both input and output variables in the first stage (DEA) can be appropriately adjusted for the evaluation model. Therefore, the inputs to the second stage are supposed to be chosen according to the expectations of insurance experts and policy makers. The end results are usually a set of rules related to both input factors and DEA productivity scores.

DEA/CART bootstrapping for evaluation of insurance branches

One of the problems of using DEA/CART is that in many DEA studies, there are not sufficient data to generate a decision tree. In view of that, the Bootstrapping technique has been proposed to increase the number of DMUs before generation of a decision tree (Emrouznejad &Anouze, 2010). This method consists of three steps. First, the values of efficiency and productivity of each branch are calculated. Then, according to the obtained efficiency and productivity values, branches will be grouped into three classes of progressive productivity (target>1, MPI>1), regressive productivity (target<1, MPI<1), and without change productivity (target=1, MPI=1).

Producing an accurate CART tree requires a large database. In case of the present study, only 18 insurance branches have been investigated; thus, by 100 times application of re-sampling bootstrapping technique, the database is enlarged sufficiently. Consequently, in the second step, 18 units (with replacement) are randomly chosen and the re sampling bootstrapping technique is applied for 100 times to obtain 1800 units. After 100 times re-sampling, the data base is divided into training and testing groups with ratio of 7 to 3.

In the third step, based on classified efficiency scores (< 1، =1، 1 >) as target variables and other uncontrollable variables (branches' rating, location, number of employees, etc.) as inputs to the CART tree, the logical decisions are extracted.

EXPERIMENTAL RESULT

DEA (first stage)

In efficiency and productivity literature, the key factors to identify input and output variables of each insurance branch are its financial balance sheet and amounts of income, profits, and losses.

Indices used in this thesis were collected over a long period of time, with reference to every branch, and based on the managers' point of views (Table 2). Then using the Latent Variable Model (LVM), efficiency and productivity of the insurance branches in the years 2008-2010 was measured.

During the review process of productivity in the years 2008-2010 , five branches displayed progression and 13 branches showed regressive trends. Similarly, in the years 2009-2010,8 branches were productive and 10 branches were not. On average, 36.2% of the branches were productive and 66.2% were not. However, due to the high dispersion, all values of the input data were

Table 2: Input/output variables in DEA

Variable ($)	Min	Max	Mean	Std
Inputs				
Administrative costs	11.944	1817808.5	179091.3	325464.4
Insurance cost	23.888	8861261.22	7146970552.3	1480265.602
No. of branches				
Outputs desirable	2	271	92.12	54.28
Revenue from insurance premiums	1085.3	12356881.9	1539634.2	2136065.4
		6168493.033		
Loan payments	6519.033	6800070.1	1941592.743	1775229.3
Output undesirable				1216054.725
Compensation payment	2756.8		1163839.4	

normalized before entering the tree for not reducing the prediction accuracy.

Bootstrapping (second stage)

As mentioned before, 18 units (with replacement) were randomly chosen and the re sampling bootstrapping technique was applied for 100 times to obtain 1800 units. This process led to a greater accuracy in the prediction of the CART tree.

CART analysis (third stage)

According to the DEA, the insurance branches were divided into three groups of progressive productivity (1< MI), regressive productivity (1>MI), and without change productivity (1=MI). These groups are used as the target variable in the CART tree.

DISCUSSION AND CONCLUSION

For all attributes, impurity levels before and after the prunings are measured and the feature that further reduced the impurities is selected. The purity index is based on the least amount of impurities in each node. In consequence, multiple regression decision trees are plotted for each period.

Regression tree analysis in the years 2008-2009

First, the prediction tree for the years 2008-2009 was considered based on variables of each branch's rating, age, location, number of employees with MA or PhD degrees, loans, issued insurance policies, and claims paid as input and productivity classification as output. Note that one of the inherent characteristics of this tree is removal of some features based on their importance or minimum correlation; thus, in this study, only

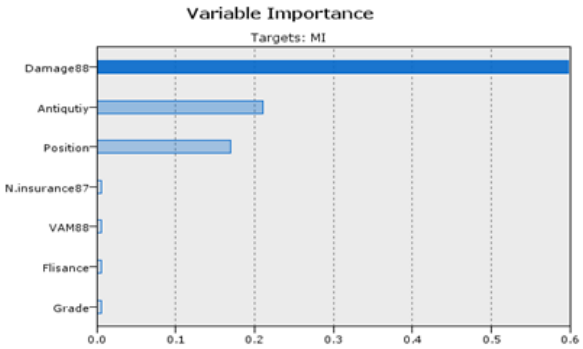

Fig.2. Importance of predictor variables 2008-2009

variables of each branch's age and location and the number of issued policies were entered into the CART tree as environmental variables.

Fig. 2. shows the importance of environmental variables. In this figure, it can be observed that the number of paid losses is the most important factor in determining the classification 59% The age (51%) and location (17%) were the second and third important factors respectively. Since the other variables were of equal importance (5%), the tree was recreated with identical variables to achieve more accurate and in-depth results. The predictive accuracy of the created tree is presented in Table 4.

As stated by the prediction, in the years 2008-2009, out of the whole 1800 cases 1290 cases had MPI< 1 and 510 cases had MPI>1. Out of 1246 training data, 1156 cases were predicted correctly with the accuracy of 92.78%

The overall accuracy of the prediction tree was 98.02% indicating a high level of confidence. Fig. 3. shows the generated CART tree with 8 nodes.

According to the presented tree in Fig. 3. the following rules can be extracted:

The following rules are extracted from insur-

Table 3: Productivity scores LVM model by Malmquist Index

Units	Of on the year	Until	MPI
DMU1	1387	1388	0.152
DMU1	1388	1389	0.267
DMU2	1387	1388	0.254
DMU2	1388	1389	0.371
DMU3	1387	1388	1.6134
DMU3	1388	1389	0.272
DMU4	1387	1388	0.0381
DMU4	1388	1389	0.471
DMU5	1387	1388	0.218
DMU5	1388	1389	0.2427
DMU6	1387	1388	0.0754
DMU6	1388	1389	1.622
DMU7	1387	1388	0.0389
DMU	1388	1389	0.252
DMU8	1387	1388	0.0502
DMU8	1388	1389	1.795
DMU9	1388	1389	0.0198
DMU9	1388	1389	0.509
DMU10	1387	1388	0.0937
DMU10	1388	1389	0.425
DMU11	1387	1388	0.003
DMU11	1388	1389	0.266
DMU12	1387	1388	0.013
DMU12	1388	1389	1.722
DMU13	1387	1388	0.001
DMU13	1388	1389	1.441
DMU14	1387	1388	0.167
DMU14	1388	1389	1.602
DMU15	1387	1388	1.882
DMU15	1388	1389	0.457
DMU16	1387	1388	1.050
DMU16	1388	1389	2.037
DMU17	1387	1388	2.663
DMU17	1388	1389	3.363
DMU18	1387	1388	2.031
DMU18	1388	1389	1.191

ance branches with progressive productivity (325 cases out of 1246 cases):

Rule 1: if the number of paid losses is smaller than or equal to 0.020, the branch has progressive productivity (192 cases).

Rule 2: if the number of paid losses is bigger than 0.020, the branch establishment year is before or in the year 2001, and it is located in west of Tehran, the branch has progressive productivity (50 Cases).

Rule 3: if the number of paid losses is bigger than 0.020 and the branch establishment year is after the year 2001, the branch has progressive productivity (72 Cases).

Table 4: Predicted accuracy of the tree

Results for output field MI
Comparing $R-MI with MI

'Partition'	1_Training		2_Testing	
Correct	1,226	98.39%	541	97.65%
Wrong	20	1.61%	13	2.35%
Total	1,246		554	

Extracting rules for insurance branches with regressive productivity

The following rules are extracted from insurance branches with regressive productivity (914 cases out of 1246 cases):

Rule 4: if the number of paid losses is bigger than or equal to 0.020, the branch establishment year is before or in the year 2001, and it is located in center, north, south, or east of Tehran, the branch has regressive productivity (912 cases).

Regression tree analysis in the years 2009-2010

Table 5 shows the predictive accuracy of the generated tree.

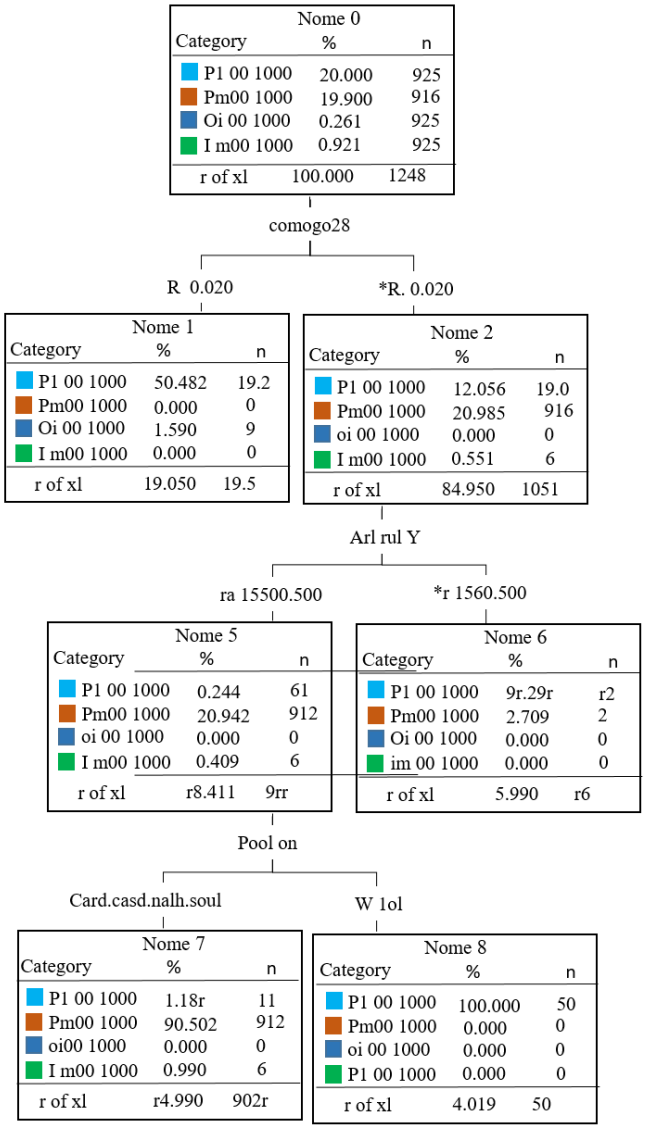

Fig.3. CART tree for Iran Insurance 2008-2009

As stated by the prediction, in the years 2009-2010, out of the whole 1800 cases 997cases had MPI<1 and 803 cases had MPI>1. Out of 1246

Table 5: Predicted accuracy of thetree.

Results for output field MI2
Comparing $R-MI2 with MI2

'Partition'	1_Training		2_Testing	
Correct	1,246	100%	554	100%
Wrong	0	0%	0	0%
Total	1,246		554	

training data, 1246 cases were predicted correctly with the accuracy of 100%.

The overall accuracy of the prediction CART tree was 100% indicating a high level of confidence.

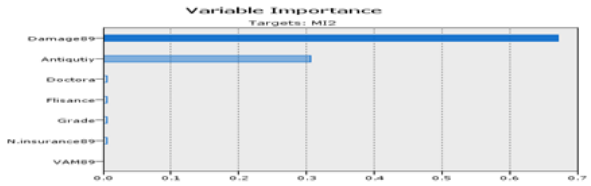

Fig.4. Importance of predictor variables 2009-2010

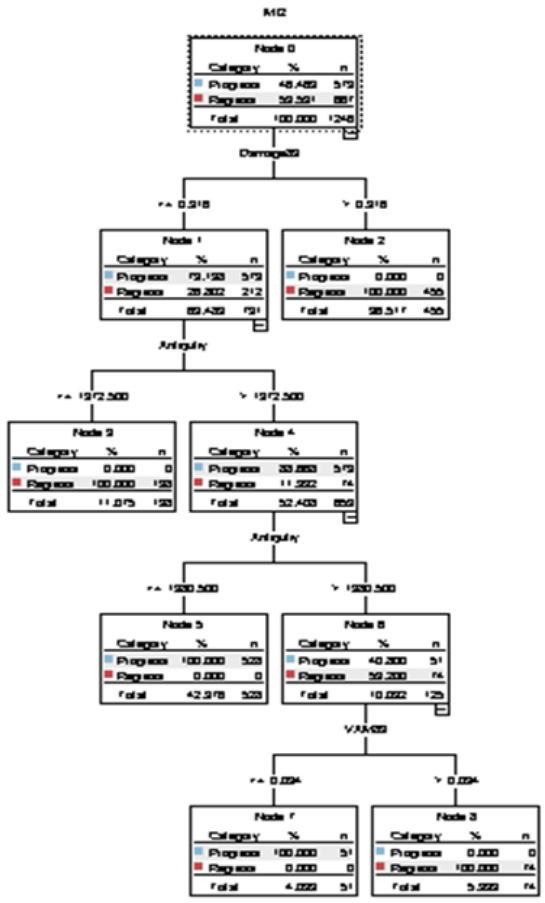

Fig.5. CART tree for Iran Insurance 2009-2010

In Fig. 5. the generated CART tree with 579 cases of progressive productivity, 667 cases of regressive productivity, and 8 nodes is presented.

In Fig .4. it can be observed that the number of paid losses in the year 2010 is the most important factor (67%). The age of the branches was the second important factor(30%) and the other variables (the number of employees with MA or PhD degrees, branch's rating, and the number of issued policies in the year 2010) were of equal importance (6%)

Extracting rules for insurance branches with progressive productivity (579 caes)

Rule 1: if the number of paid losses is smaller than or equal to 0.316 and the branch establishment year is between the years1993 and 2001, the branch has progressive productivity (528 cases).

Rule 2: if the number of paid losses is smaller than or equal to 0.316,the branch establishment year is after the year 2001, and the number of paid loans is less than or equal to 0.034, the branch has progressive productivity (51cases).

Extracting rules for insurance branches with regressive productivity (667 cases)

Rule 3: if the number of paid losses in the year 2010 is smaller than or equal to 0.316 and the branch establishment year is before the year 1993, the branch has regressive productivity (138 cases).

Rule 4: if the number of paid losses in the year 2010 is smaller than or equal to 0.316, the branch establishment year is after the year 1993, and the number of paid loans is more than 0.034, the branch has regressive productivity (74 cases).

Rule 5: if the number of paid losses is bigger than 0.316, the branch has regressive productivity (455 cases).

FINAL EVALUATION

This thesis tries to introduce a combination of DEA and CART tree. In this study, insurance branches in Tehran are examined. In general, the efficiency and productivity scores can be obtained using DEA and MPI. However, these methods cannot explain the related factors to inefficiency and unproductively, especially in case of variables that are not numerical.

Considering factors associated with efficiency and productivity, CART tree can present a better understanding of the DEA results. Despite the proposed method in the present study is examined in the insurance industry, it potentially has much

broader applications. Regarding DMUs'efficiency and productivity evaluation, the proposed DEA/CART method can be applied as a framework for further research.

The results of this combined method are a set of rules, which can be applied by policy makers to explore the reasons behind DMUs' efficiency and inefficiency. Creating a good and reliable CART tree usually requires a large database and many observations; but in most of the reported DEAs, the number of DMUs is not large enough to generate a proper CART tree. In order to solve this problem, the Bootstrapping method was proposed in this study. Nonetheless, further investigations seem quite necessary for an appropriate application of this method.

CONCLUSION

Data Envelopment Analysis (DEA) is a management tool for efficiency and productivity assessment. This paper presented a framework for relating DEA to classification and regression analysis. While the DEA provides valuable and acceptable results, the CART analysis reveals additional facts that were unclear in previous studies.

Unlike previous studies in the fields of DEA and insurance industry that just tried to identify the impacts of different factors on the efficiency with the same impact level, the proposed CART tree is based on the analysis of impact levels of different factors related to efficiency and productivity of insurance branches.

Exploring the variables' importance and influence on variables' dependence with the least amount of impurities to reach the target node (through the Clementine software), can lead to an in-depth analysis with the lowest amount of error by combining environmental factors with efficiency and productivity scores (obtained from the DEA).

In previous studies, only the key parameters in the efficiency or inefficiency of the insurance branches have been evaluated and no environmental factor related to progressive/regressive productivity has been addressed yet. For example, the number of losses, paid loans, and age of the branches are not considered as important factors in the efficiency/inefficiency issue; however, according to the extracted rules, they are influential variables in the efficiency/inefficiency evaluation of the insurance branches with different impact levels.

Furthermore, using numerical and categorical variables with different degrees of importance, rules were extracted for each specific DMU and used to identify productivity or unproductively of the selected insurance branches.

Unlike previous studies on DEA applications, which focused only on the numerical calculations of efficiency and productivity, this paper studied factors related to efficiency and productivity of insurance branches, using CART tree. In addition, possible rules were extracted for every DMU, using both numerical and categorical variables. Obviously these rules are very useful for policy makers and can improve their decision-making processes.

FUTURE STUDIES

There are a number of additional issues of practical importance to those who study CART trees (independent factors for the insurance sectors, application of various rules and accurate measurement, and improvement of the Bootstrapping method). Despite these issues have not been addressed in the current investigation, their inclusion in other studies can broaden the field for the development of future studies.In future research, databases with larger sample size can be chosen to avoid using the Bootstrapping method. It must be noted that the use of simulation in this paper was one of the limitations.

Fuzzy decision tree can be used instead of crisp decision tree because it offers beneficial results in case of insurance industry's qualitative data. It is also possible to set the DEA efficiency and productivity results as output variables. Moreover, depending on the type of data and the importance of input variables, other trees such as, and can be used.

REFERENCE

Barr, R., SeifordL, M., & Siems, T.F. (1994). Forecastingbank failure: a non-parametric approach. *Recherches Economies de Louvain, 60*,411-429.

Bretholt, A., & Pan, J. N. (2013). Evolving the latent variable model as an environmental DEA technology. *Omega, 41*(2), 315-325.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C.J. (1984). Classification and regression trees. Wadsworth & Brooks. *Monterey, CA*.

Clark L., Pregibon, D. (1992). Tree based models, in J.M. Chambers and T.J. Hatie (eds), Statistical Models in S, Pacific Grove, CA: Wadsworth

& Brooks/Cole Advanced Books & Software, 377-419.
Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring the efficiency of decision making units. *European journal of operational research, 2*(6), 429-444.
Chen, Y.L., Hsu, C.L., & Chou, S.C. (2003). Constructing a multi-valued and multi-labeled decision tree. *Expert Systems with Applications, 25*(2), 199-209.
Emrouznejad, A., & Anouze, A. L. (2010). Data envelopment analysis with classification and regression tree–a case of banking efficiency. *Expert Systems, 27*(4), 231-246.
Färe, R., & Grosskopf, S. (2008). A comment on weak disposability in nonparametric production analysis. *American Journal of Agricultural Economics, 91*(2), 535-538.
Han, J., & Kamber, M. (2001). Data Mining Concepts and Techniques, San Francisco, CA: Morgan Kaufmann.
Hand, D.J., Mannila, H., & Smyth, P. (2001). Principles of Data, Cambridge, MA: MIT Press.
Lee, J.H., & Park, S. C. (2005). Intelligent profitable customers segmentation system based on business intelligence tools. *Expert systems with applications, 29*(1), 145-152.
Malmquist, S. (1953). Index numbers and indifference surfaces. *Trabajos de estadística, 4*(2), 209-242.
Pille, P., & Paradi, J. (1997). Facets at the frontier and efficiency measurement in DEA, Paper presented at the Fifth European Workshop on Efficiency and Productivity Analysis, Copenhagen, October.
Ripley, E.M., & Miller, J.D. (1996). Layered intrusions of the Duluth complex, Minnesota, USA. *Developments in Petrology, 15*, 257-301.
Samoilenko, S., & Osei-Bryson, K.M. (2013). Using Data Envelopment Analysis (DEA) for monitoring efficiency-based performance of productivity-driven organizations: Design and implementation of a decision support system. *Omega, 41*(1), 131-142.
Seol, H., Choi, J., Park, G., & Park, Y. (2007). A framework for benchmarking service process using data envelopment analysis and decision tree. *Expert Systems with Applications, 32*(2), 432-440.
Sohn, S. Y., & Moon, T. H. (2004). Decision tree based on data envelopment analysis for effective technology commercialization. *Expert Systems with Applications, 26*(2), 279-284.
Zhu, J. (2004). Quantitative models for performance evaluation and benchmarking. Boston: Kluwer Academic Publishers.