Indian Journal of Fundamental and Applied Life Sciences ISSN: 2231– 6345 (Online)
An Open Access, Online International Journal Available at www.cibtech.org/sp.ed/jls/2015/02/jls.htm
2015 Vol. 5 (S2), pp. 1795-1803/Mohammadzade and Mohammadjafari
Research Article

ANALYZING FAILURE MODE USING DATA MINING AND DATA ENVELOPMENT ANALYSIS

Mohammad Mohammadzade and *Marjan Mohammadjafari
Department of Industrial Engineering, Kerman Branch, Islamic Azad University, Kerman, Iran
**Author for Correspondence*

ABSTRACT
By occurring industrial revolution and developing production issues in today's world sciences, repair and maintenance programming has got double importance. Many of industrial centers and factories do considerably pay attention to repair and maintenance costs and they are doing their best to use preventive schedules to avoid from production line shut down. By applying this action, they optimize their repair costs. One of this kind of industrial centers is refineries, because their production line stoppage costs is very high. One of the most prevalent techniques applied in failure mode and effect analysis is Risk priority number (RPN) which considers 3 parameters of severity, occurrence and detection and it ranks failure modes based on 3 aforementioned criteria and 2 other criteria (cost and importance of each failure mode). But since, number of failure modes is high, using this approach can't be solely appropriate. Techniques which analyze massive databases have been developed entitled "data mining". In this paper, to reduce calculations and enhance results' accuracy, data mining will be used for data clustering. On the other hand, due to this fact that conducted clustering can't be only used, multi criteria decision making is employed in order to rank several alternatives based on similar criteria. Various and different decision making techniques are used by researchers, but one of the most applicable one is data envelopment analysis which is based on efficiency rate of each of units based on rate of output to input. This paper has been conducted over data from Tehran refinery and results obtained from this paper can be considerably helpful for correct ranking and clustering failure modes in aforementioned refinery.

Keywords: *Data Mining, Multi Criteria Decision Making, Data Envelopment Analysis, Failure Mode and Effect Analysis, Ranking, Repair and Maintenance Programming*

INTRODUCTION
Modern attitudes related to quality guarantee requirements, customers' increasing expectations and sever competition in the arena of international commercial interactions has encouraged producers to take plenty of efforts in the field of quality enhancement and development.
In this case, one of the newest methods which have been proposed in industrial countries is FMEA or failure mode and effect analysis.
This technique has been firstly applied in US aerospace industry in 60s, but its fast growth and development was occurred in 70s and 80s in automotive industry and today it is considered as one of requirements for automotive producers.
FMEA is in fact a kind of method for analyzing potential failures 'risk which engineers are encountering while designing and producing a new production. This technique systematically identifies potential and possible failures. Coefficients of risk priority number (RPN) are measuring all of failures by combining factors such as failure severity, failure occurrence and failure detection and then proposes a priority to reduce risks existed in designing and producing new production (Stamatis, 1995).

Literature Review
Soleimani *et al.,* (2010) in a paper entitled "Performance evaluation of clustering algorithms", by choosing some of clustering algorithms in data mining and implementing them over more than 50 databases using DEA, have ranked algorithms.
Bamakan and Gholami (2011) have proposed a paper entitled "FMEA using Fuzzy method and Grey theory". In this paper, conventional model of FMEA was modified and modeled in a fuzzy form via considering 3 parameters of failure severity, failure occurrence and failure detection and a method was

proposed to prioritize failures and their effects in a fuzzy environment and also by applying Grey theory and allocating weight ratio to each of these three parameters, relationship between them was considered.
Rad *et al.*, (2011) have taken an action to rank and cluster Iran universities main fields in bachelor degree. Accordingly, fields were clustered based on their similarities in 10 main clusters using K-MEAN algorithm (Rad *et al.*, 2011).

Problem Statement

Increasing expansion of industrial products in order to supply human daily needs has caused that repair and maintenance issue has become one of the most important issues in industrial firms. Because, available resources are getting limited and limited and achieving these resources is undoubtedly dependent on correct use of today's advanced industries. That's why; applying optimal techniques in analyzing and detecting faults of equipment has been significantly addressed.
Since, all of human-made machineries have a limited life, it is better to think about a strategy by optimizing machinery lifetime which has understandably constituted of various components and in each time, there is a possibility of failure of each component and subsequently system breakdown.
Since, repair and maintenance costs is one of important cost items in products' net price, accuracy in this context can smooth a way to retain survival in competition era. That's why; knowing appropriate patterns and prioritizing repair and maintenance activities in this path will be helpful. In order to maximize efficiency and equipment availability, there is a need for accurate and continuous programming. In this case, gaps which sometimes are leading to high costs in production are reduced by predicting and resolving defects or components problems.
FMEA technique is an analytical technique based on the law called "prevention before occurrence" which is used to identify potential failure factors. In failure detection of equipment, it is an analytical method which is trying to investigate current potential dangers and also identifies and ranks effects related to that. This technique is used to calculate risk number. Indeed, RPN shows the rate of risk taking by failure factors which is obtained from multiplying 3 parameters of failure severity, failure occurrence and failure detection.
But, in studying a complicated system having diverse machines, need for prioritizing failure rate and effects derived from that is a very important issue. That's why, in order to determine priorities of failures handling, importance of applying K-Mean technique is quite transparent. K-mean algorithm is one of most important and applicable clustering algorithms. In this algorithm, samples are divided into K clusters, so due to this issue that studying all of failure possible cases is so time-taking, in order to determine underlying and important activities and conduct FMEA technique over equipment, clustering is used. Ultimately, in this dissertation, data envelopment analysis will be employed to enhance clustering results and more appropriate programming in order to conduct recommended method.

MATERIALS AND METHODS

Multi Criteria Decision Making

Multi criteria decision making is one of the decision making methods to choose alternatives among a set of solutions. Many of multi criteria decision makings are making efforts to draw an ideal decision making environment in which, decision makers are logically addressing all of problem's aspects and obtaining all of accurate information and then accept one solution by consensus. Multi criteria decision making is the best-known branch of decision making. This branch is one of the general classes of research in operation models which has been initiated to solve decision problems which have plenty of decision criteria (Ong and Yon, 2008).
This excellent class of models has been concentrated on the problems with discontinuous decision space. In these problems, a set of decision alternatives has been previously determined. Although, multi criteria decision making techniques have been massively diversified, many of them have the common aspects. In addition, each model also has its own features.
The fundamental of MADM models consists of 2 underlying steps (Gholami *et al.*, 2011):

1- Cumulating judgments due to each alternative and criteria.

2- Criteria ranking based on cumulative rules.

Two main categories of different methods in processing information obtained from one MADM problem in literature review have been proposed. One category of methods is originated from a famous model called Non-compensatory model and another category is derived from another model known as Compensatory model.

a) Non-compensatory models: non-compensatory models include methods in which Trade-off is not permitted among indices. That is, weak point existed in one index is not compensated by another advantage existed in another index. Hence, each index in these methods is solely proposed and comparisons are done based on index to index. This model consists of techniques which mostly don't need to obtain information from DM and they will lead to one objective solution. Application of these relatively simple methods is dependent on decision making status and they demand more accuracy by analyst. These methods are as the following:

1- Dominance method
2- Maxim in method
3- Conjunctive-satisfying method
4- Disjunctive-satisfying method
5- Lexicography method
6- Elimination method
7- Permutation method

Advantage of methods belonging to this model is their simplicity which is consistent with DM behavior and information limitation. In some of methods, it might need to gain information from DM.

b) Compensatory models

This model consists of methods in which trade-off among indices is not legible. That is, for example a change (probably trivial) in one index can be compensated by an opposite change in index (or other indices). This model is divided into 3 sub-groups as the following:

a- Privileging and scoring subgroup
b- Compromise subgroup
c- Coordinator subgroup (Pang *et al.*, 2011)

Data Envelopment Analysis

Data envelopment analysis is a nonparametric method which determines efficient border of those kinds of decision making units (DMU) which have similar inputs and outputs.

In DEA method, there is no need for any kind of hypothesis or an especial mathematic form. That is, there is no need for knowing production function. Also, having prices of production factors is not necessary. Therefore, in a circumstances in which existed information is not sort of information which can't be estimated by applying production function or there is no price of production factors, DEA is an appropriate method to measure efficiency. Obtained efficiency in DEA method is a relative efficiency and efficiency border is created by a convex combination of efficient units. So, each enterprise which is located over the efficiency border is efficient and otherwise, it is inefficient. In order to make an efficient unit, some changes must be happened in its inputs and outputs. It can be mentioned that after DEA models, a set so called "Reference collection" is identified. In this set, it is determined that each inefficient unit has to be compared with which one of efficient units in order to achieve efficiency frontier (Charnz *et al.*, 1984).

Although, DEA models are increasing and also they are finding specialization aspect, but fundamental of all is number of main models which have been designed and expressed by the founders of this methods.

Initial Model of Input-oriented CCR

In order to transform CCR model to a linear programming model by the method used by Charnez and Cooper and in order to maximize the value of deduction phrase, it is enough to consider denominator equal to a constant figure and maximize nominator. Accordingly, denominator is considered equal to one number and a new model is obtained as the following. This model is called initial form (multiple) of CCR (Mehregan, 2004).

$$Max Z_o = \sum_{r=1}^{s} u_r y_{ro}$$

$$s.t:$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0$$

$$\sum_{i=1}^{m} v_i x_{io} = 1$$

$$u_r \ge 0$$

$$v_i \ge 0$$

Model*(1)*

(j = 1,2,3,...,n) و (r= 1,2,3,...,s) و (i=1,2,3,....,m)

Data Mining

Data mining is one of the fastest growth contexts for computer industry. When an interesting specific and restricted context in statistic and computer field is opened, it is rapidly improved and developed to its real place and level. One of the most important abilities of data mining is appeared in the wide range of methodologies and techniques which can be applied to host a set of different problems. According to this issue that data mining is a natural activity which is applied over a massive and voluminous data set, one of the largest target markets, data comprehensible stores, specialized data centers and decision support system is to gain specialties in industries such as retailing, production, telecommunication, public sanity, insurance and transportation (Yap *et al.*, 2011). In commercial issues, data mining is employed to propose new purchase methods, investment strategies and detecting illegal costs in accounting system. Data mining can enhance competition and marketing output and attract customers' earnings, support and satisfaction. Also, data mining techniques can be used to solve commerce process. In fact, the main purpose is to perceive interaction and connection between business methods and required organizations. Data mining duties are divided into 4 main categories. One of these duties is data clustering in the various clusters with respect to similarity of each cluster. Its goal is to optimize distance of each elements of cluster from cluster's center. This task is done by plenty of algorithms, but the most important and popular algorithm is K-means algorithm (Raiten and Frank, 2011).

K-means Algorithm

This algorithm considers K parameter as an input and it partitions a set of n objects to K cluster. So that, the level of clusters' internal similarity is high and similarity level of objects out of clusters is low. Similarity of each cluster is measured compared with average of objects in that cluster which is called cluster center. This algorithm is working as the following:

Input: K is the number of clusters and database includes n objects.

Output: A set of k clusters which minimize square-error criterion.

Algorithm:

Step 1: Select k arbitrary points randomly as the center of initial clusters (it is better k points is chosen from n current points).

Step 2: Allocate each object to the clusters based on its similarity to the clusters' center.

Step 3: Update clusters' center. That is, calculate mean of objects of each cluster for each cluster.

Step 4: Back to the second step according to clusters' new center until no change is made in clusters (In this case, algorithm is finished).

In action, this algorithm is a heuristic method for reducing square-error criterion which has been brought here.

$$E = \sum \sum |p - m_i|^2$$

In the relation, E is the sum of square-errors for all of objects in database. P is a point in the space which is exhibiting one object and m_i is the mean of cluster C_i which point p belongs to that (either p or m_i are multidimensional).

This algorithm is appropriately working when clusters are like intensive clouds and these clouds are separated per se. This method is not efficient for huge database and has to be extended. Its calculative complexity is O(tkn) that:

N is the number of objects. K is the number of clusters and t is number of algorithm repetitions. This method is mostly terminated to a local optimum not a worldwide optimal.

K-mean technique is applied only when clusters' center can be defined. For example, this technique is not efficient for data with the class features.

One of the disadvantages of this method is to determine k which must be firstly determined by user and there is not a specific way to determine that.

One way is to test different Ks and estimate square-error criterion for each K. Also, this method is not proper to discover clusters with complicated forms.

One of the most important weak points of this method is that it is sensitive toward disturbances and outliers, because these data are easily changing centers and desirable results might be not obtained.

To resolve defects of K-means algorithm, some changes happened over that.

These extended methods are different with each other in selecting k initial centers, calculation of dissimilarity and strategies of calculating clusters' centers.

One of these changes is that at first, it is implemented over a hierarchy cumulative algorithm in order to find number of desirable clusters and then obtained clusters are used as the first step of K-means algorithm (Rad and Naderi, 2011).

Failure Mode and Effect Analysis

Failure mode and effect analysis (FMEA) is widely using engineering techniques to define, identify and eliminate identified cases, failures' potential, problems, mistakes and so on in system, design, processes and services before proposing customer (Stamatis, 1995). Failure mode and effect analysis is defined as a method in components, system and processes.

A failure in one component can cause failure in other components. The reason behind a failure is defined as a weakness of design and it might lead to more general failures. For each failure, potential failure mode is identified which is ultimately emerged.

It needs to be determined by FMEA team. Failure effect is a result of potential failure in function, production and process in the viewpoint of customer.

One system, design, process or service usually has several failures modes, effects and reasons. Under this circumstance, each potential failure and its effect need valuing and prioritizing based on loss possibility. Failure mode and effect analysis is able to identify priorities. In traditional FMEA, prioritizing failure mode is calculated via risk priority number (RPN) which is derived from multiplying failures' severity, occurrence and detection. That is,

$$\mathrm{RPN} = \mathrm{O} \times S \times D$$

Where O and S are occurrence multiplicity and severity of failures' effects respectively and D is occurrence before delivering customer. All 3 risk factors are measured by 10 scales from 1 to 10. Potential failure with the higher RPN is considered as more important mode (Laura, 2009).

MATERIALS AND METHODS

Research Methodology

Research Analytical Model

The main model in this dissertation is according to chart 1.

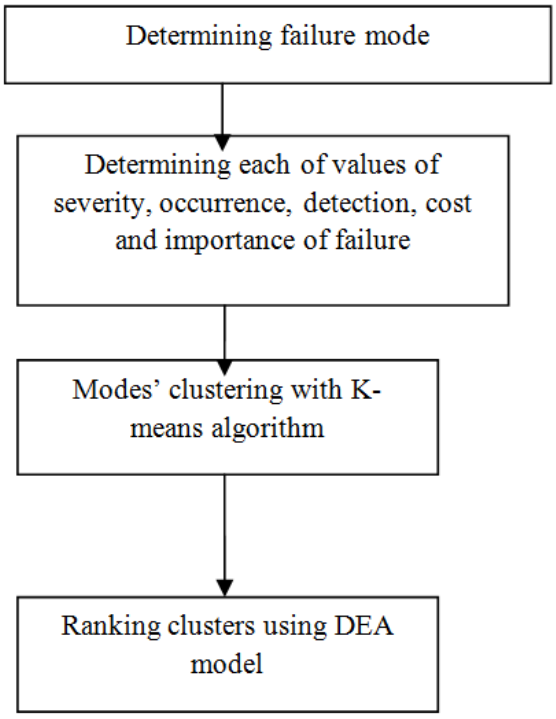

Chart 1: Research analytical model

Case Study

Establishing Database

At first, a database must be set up. Accordingly, it is done using tables existed in repair and maintenance sector.

With respect to current tables, 1035 points have been chosen as the failure modes. In addition to the values of severity, occurrence and detection, experts have recommended 2 other criteria as criteria affecting on analyzing points. These 2 criteria also have been previously valued for all points which are failure cost (it is positive kind) and impotence of failure point in the sector performance (it is positive kind).

Modes' Clustering

Based on aforementioned analytical model in this paper, we are approaching to cluster points in which data are divided into 15 clusters. Clusters' centers in each of criteria have been displayed in table 1.

Table 1: Coordinate of clusters' centers

	Clusters														
Criteria	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Severity	4.00	2.00	10.00	1.00	1.00	9.00	8.00	8.00	3.00	9.00	8.00	10.00	10.00	2.00	1.00
Occurrence	2.00	1.00	3.00	9.00	7.00	1.00	1.00	10.00	9.00	1.00	9.00	6.00	10.00	1.00	9.00
Detection	10.00	1.00	3.00	7.00	1.00	10.00	1.00	7.00	5.00	8.00	1.00	1.00	10.00	10.00	10.00
Cost	10.00	10.00	9.00	9.00	1.00	10.00	1.00	2.00	9.00	2.00	1.00	9.00	9.00	3.00	2.00
Importance	9.00	5.00	9.00	1.00	4.00	1.00	2.00	1.00	10.00	9.00	10.00	1.00	7.00	3.00	6.00

Clusters' Ranking using DEA

First, we are approaching to set up a model in which alternatives are 15 cluster's center and also outputs are severity, occurrence, detection, costs as well as importance of clusters' centers based on table 1 and input about each of clusters is constant number 1.

In this step, data envelopment analysis model is built and then it is solved. Efficiency of each cluster has been exhibited in table 2.

Table 2: Efficiency of each cluster

Cluster	Efficiency
1	1
2	1
3	1
4	0.99
5	0.7
6	1
7	0.8
8	1
9	1
10	1
11	1
12	1
13	1
14	1
15	1

Since, Clusters'' efficiency is 1 in most f cases, so, Anderson-Peterson model has been used to determine efficiency and clusters' final ranking which have been shown in table 3.

Table 3: Final efficiency and clusters' ranking

Cluster	AP efficiency	Rank
1	1.15	4
2	1	7
3	1	7
4	0.99	13
5	0.7	15
6	1.08	5
7	0.8	14
8	1	7
9	1.21	2
10	1.09	6
11	1.16	3
12	1	7
13	1.46	1
14	1	7
15	1	7

As an example, efficiency of cluster 1 has been determined from the following programing.

Max=4*x + 2*y + 10*z + 10*u + 9*v;
2*x + 1*y + 1*z + 10*u + 5*v=<1;
10*x + 3*y + 3*z + 9*u + 5*v=<1;
1*x + 9*y + 7*z + 9*u + 1*v=<1;
1*x + 7*y + 1*z + 1*u + 4*v=<1;
9*x + 1*y + 10*z + 10*u + 1*v=<1;
8*x + 1*y + 1*z + 1*u + 2*v=<1;

8*x + 10*y + 7*z + 2*u + 1*v=<1;
3*x + 9*y + 5*z + 9*u + 10*v=<1;
9*x + 1*y + 8*z + 2*u + 9*v=<1;
8*x + 9*y + 1*z + 1*u + 10*v=<1;
10*x + 6*y + 1*z + 9*u + 1*v=<1;
10*x + 10*y + 10*z + 9*u + 7*v=<1;
2*x + 1*y + 10*z + 3*u + 3*v=<1;
1*x + 9*y + 10*z + 2*u + 6*v=<1;
x>=0;
y>=0;
z>=0;
v>=0;

RESULTS AND DISCUSSION

Due to importance of repair costs for the firm and attempts done by repair and maintenance units in regard with reducing these costs, proposing preventive schedules is very important and this preventive schedule must have the capability of high flexibility.

In this case, experts can take an action to cluster and program in order to prevent from failure in these points based on their respective needs and criteria. Also, in clusters 'final evaluation, a technique has to be used which is doing ranking with the higher accuracy. Hence, data envelopment analysis technique seems suitable to rank clusters.

In regard with enhancement of results in this research, it is recommended to other researchers to use fuzzy logic to determine centers for each of clusters and also clusters' virtual clusters. Also, weight of factors affecting on ranking is obtained using hierarchy analysis and network analysis and also, they can determine other factors affecting on ranking using other decision making techniques and modern approaches.

Conclusion

According to the results obtained from this paper, failure modes in Tehran refinery are categorized in 15 main clusters. If experts want to consider criteria of severity, occurrence, detection, cost as well as importance of each failure in analysis and programming, strict preventive maintenance schedule should be applied for the points which are located in cluster 13 and then it is used for the points located in cluster 9 and afterward, cluster 11 has to be addressed and similarly, it has to be considered for clusters 5, 7 and 4 which have the least importance.

Table 4: Conclusion

Low priority clusters	High priority clusters	Applied criteria	Mode
5-7-4	13-9-11	Severity, occurrence, detection, cost and importance	4

REFERENCES

Gholami P, Bazleh A and Salehi M (2011). Using Clustering and DEA for evaluation and ranking university majors. *Journal of Computing* **3**(8) 18.

Hwangn CL and Yoon K (2011). *Multiple Attribute Decision Making*, (2008) (Springer Verlag).

Maria Laura Chiozza and Clemente Ponzetti (2009). FMEA: A model for reducing medical errors.

Peng Y, Zhang Y, Tang Y and Shiming L (2011). An incident information management framework based on data integration, data mining, and multi-criteria decision making. *Decision Support Systems* **51** 316–327.

Rad A, Naderi B and Soltani M (2011). Clustering and ranking university majors using data mining and AHP algorithms: A case study in Iran. *Expert Systems with Applications* **38** 755–763.

Shi Y, Peng Y, Kou G and Chen Z (2005). Classifying credit card accounts for business intelligence and decision making: A multiple-criteria quadratic approach. *International Journal of Information Technology & Decision Making* **4**(4) 581-599.
Stamatis DH (1995). *Failure Mode and Effect Analysis: FMEA from Theory to Execution* (Milwaukee, WI: ASQC Quality Press).
Witten Frank IH (2011). *Data Mining—Practical Machine Learning Tools and Techniques*, second edition (San Francisco, CA: Morgan Kaufman).
Yap BW, Ong SH and Husain NHM (2011). Using data mining to improve assessment of credit worthiness via credit scoring models. *Expert Systems with Applications* **38** 13274–13283.