MINING EMPLOYEE EFFICIENCY IN MANUFACTURING

ESRA Kahya Ozyirmidokuz[1], CEBRAIL Ciflikli[2]

Erciyes University

Abstract

Data mining has a significant impact in information management in discovering patterns in order to use the gained knowledge to develop strategies and future plans. Employees are one of the determining factors for the success of a firm in a global economy. We use data mining to discover useful patterns from the employee faults in manufacturing to enhance employee efficiency. Employees can understand the "why" behind their jobs with the help of the generated rules to be successful in the job. The acquired decision tree model is different from existing methods that have problems such as difficulty to understand.

Keywords *Information systems, Decision analysis, Data mining*

JEL classification: *C44, C53*

1. Introduction

Measurement plays an important role in an organization to improve productivity. It helps to determine the organization's situation. In addition, measuring employee performance is a key strategy for a firm's success. Managers must determine where inefficiencies exist and identify strong employees by detecting the negative/positive trend among a number of employees.

DM, which is a powerful tool for data analysis, improves the competitiveness of the firms and the economic condition of employees in the

[1] *Assistant professor / Ph.D., Erciyes University Computer Technologies Department, esrakahya@erciyes.edu.tr*

[2] *Professor Dr. / Ph.D, Erciyes University Electronical and Automation Department, cebrailc@erciyes.edu.tr*

sector of the economy. We explore employee efficiency for predicting manufacturing efficiency of the employees.

As the importance of the knowledge-based economy increases, DM is becoming an integral part of business and governments. Many applications have been incorporated into the information systems and business processes of companies in a wide range of industries. Although less publicized, DM is becoming equally important in science and engineering. (Soares and Ghani, 2010). Having the right information at the right time is crucial for making the right decision. The problem of collecting data, which used to be a major concern for most organizations, is almost resolved. In the millennium, organizations will compete in generating information from data and not in collecting data. Industry surveys indicated that over 80 percent of the companies listed in Fortune 500 believe that DM would be a critical factor for business success by the year 2000. Obviously, DM will be one of the main competitive focuses of organizations. Although progress is continuously been made in the DM field, many issues remain to be resolved and much research still needs be done. (Lee and Siau, 2001).

The DM approach in human resources (HR) involves the analysis of data in large databases, and has become a useful tool for the HR professional by extracting knowledge based on patterns of data from large databases. It not only performs data analysis on a large dataset, but also provides the compony with a competitive advantage, by managing HR resources in an organization, and can bridge the knowledge gap (Alsultanny 2013).

Nowadays, the importance of DM, which is an interdisciplinary analysis, is in the process of emerging. With the increase in the popularity of big data, DM techniques have been used in the field of HR. DM provides information on how effectively and efficiently an organization manages its resources. Employee efficiency can be determined by using DM techniques. This study aims to detect employee efficiency via DM. It also aims to examine the effects of employee fault factors. We use the CRISP-DM (Cross Industry Standard Process- DM) process model. It contributes to DM as a process which is reflected in its origins (Colleen, 2006). According to CRISP-DM, a given DM Project has a life cycle consisting of six phases (business understanding, data understanding, data prepration, modeling, evaluation and deployment), as illustrated in Fig 1 (Larose, 2005).

We generate a decision tree (DT) model to to simplify the decision making process. A DT represents the decisions in a human-understandable

structure. There are numerous advantages of DTs in many classification and prediction applications. The DT is an easy model to interpret and explain to managers. Managers can see all possible outcomes from a DT. In addition, nonlinear relationships between attributes do not affect the DT's performance. DTs also cover the forecasting of future outcomes.

The paper is organized as follows. The literature is presented in Section **2**. Data are described in Section **3**. Section **4** presents the data understanding phase. The preprocessing techniques are applied in Section **4**. A DT based DM model is developed in Section **5**. In Section **6**, we draw our conclusions.

2. Literature

The use of DM techniques in manufacturing began in the 1990s. In recent years, the literature has presented several studies that examine the implementation of DM techniques in manufacturing (Ciflikli and Kahya-Özyirmidokuz, 2008; Ciflikli and Kahya-Özyirmidokuz, 2010; Ciflikli and Kahya-Özyirmidokuz, 2012). Wang (2007) discussed the nature and implications of DM techniques in manufacturing and their implementations on product design and manufacturing. Harding et al. (2006) reviewed applications of DM in manufacturing engineering, in particular production processes, operations, fault detection, maintenance, decision support, and product quality improvement.

In recent years, a rapidly growing number of research contributions have aimed at supporting the practical adoption of HR management DM (Strohmeier and Piazza 2013). Ramesh (2001) proposed an inductive DM technique (named GPR) based on genetic programming. Using a personnel database of 12,787 employees with 35 descriptive variables, Ramesh discovered employees' hidden decision making patterns in the form of production rules. Thissen-Roe (2005) developed a hybrid system which captures the main advantages of both technologies: the modeling flexibility of a neural network, and the efficiency gains of adaptive testing. A prototype was implemented for the case of personality assessment used to predict job tenure in a national retail chain. Zhao (2008) introduced a new method of DM that can be applied in the performance evaluation of HR management. It is aimed at discovering the relationship among modes and data by all kinds of analysis methods. He chose the performance evaluation of HR management of in medium-sized company as the research objective.

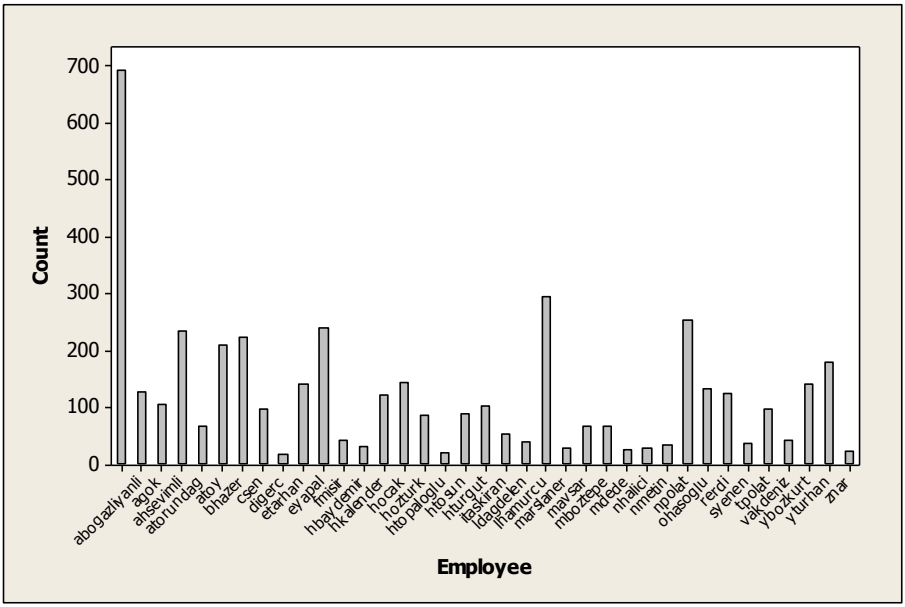
Karahoca et al. (2008) presented a self-regulating clustering algorithm for identifying a suitable cluster configuration without a priori knowledge of the given data set. An online self-regulating clustering method was applied to the HR performance data, which were gathered from 1,100 employees' yearly performance scores for two years. Li et al. (2009) proposed a new method based on support vector machines in order to solve the problem of large sample and low resolution problems in predicting the risk of HR in construction enterprises. Aiolli et al. (2009) proposed a method that incrementally builds a committee of classifiers (experts), each one trained on the newer chunks of samples. They tested their approach on a large dataset coming from many years of HR selections in a bank. Juan (2009) used the data flow of state-employed professional and technical personnel for the sample to show the stability of state-owned enterprises' human capital from 2001 to 2007 via neural networks. Sivaram and Ramar (2010) applied k-means and fuzzy c-means clustering and DT classification algorithms to the recruitment data. Experiments were conducted with the data collected from an information technology company to support their hiring decisions. They generated ID3, C4.5 and CART DTs in a comparative DM research. Aviad and Roy (2011) proposed a new technique to define DT based on cluster analysis. The proposed model was applied and tested on two large datasets of real life HR classification problems. Lockamy and Service (2011) focused on managerial and professional career advancement research, managerial promotion processes, and personnel development. Factor analysis was used to determine the most influential factors, and Bayesian networks were constructed to determine the probability of receiving a promotion based on these factors. Zhang and Deng (2011) used association rules on HR management in a university. They used the data in fact tables of one year, applied the betterment association rule mining algorithms, mining the relationship among the basic properties and their mutual influence, such as teacher's age, title, degree, gender, etc. Alsultanny (2013) used Naïve Bayes, DTs and decision rules to classify unknown instances for employment automatically. Strohmeier and Piazza (2013) reviewed DM research on HR to systematically uncover recent advancements and suggest areas for future work.

This research is a continuation of a previously presented study (Ciflikli and Kahya Ozyirmidokuz, 2008). Unlike different from other studies, we mine employee fault data in manufacturing and we extract knowledge to understand the efficiency performance of the employee via DTs.

3. Data preparation phase

The first step of the knowledge discovery process is to define what the problem really is. When a problem description is found, according to the definition, then a process of planning is initiated. After understanding the whole set of data and what can be extracted from it, the objective of the study is determined. The analysis of the subject is detailed, followed by the understanding and preparation of data (Çiflikli and Kahya-Özyirmidokuz, 2012). Data are collected from a carpet manufacturing factory in Turkey and 4,431 samples are used. The attributes are as follows: fault reason, month, shift, workbench, employee, quality, design, color, width, height, fault type and department. The employee with the most faults can be seen from Figure 1. This does not mean he is the most inefficient employee. Other attributes are also effective in the formation of an employee's fault. In addition, the maximum number of employee faults is in August.

Figure 1: Chart of employees



There are 10 workbenches and 3 shifts in the process, which has 9 departments. There are 12 size variables, which show the size of the carpet. Shift, height and width are ordered set typed integers. There are 24 quality and

18 color codes for carpets. Design attribute indicates the 200 design codes of the carpets in the process. There are 3 variables of fault type attribute in the process; HA (faulty), IF (extra manufacture) and AZ (low faulty). Besides these variables, there are five more categorical choices because of the data incompatibility. The fault reason attribute has 97 variables in the data matrix.

After the data have been examined and characterized in a preliminary fashion during the data understanding stage, the data are then prepared for subsequent mining and analysis. This data preparation includes any cleaning and recording as well as the selection of any necessary training and test samples. It is also during this stage that any necessary merging or aggregating of data sets or elements is done. The goal of this step is the creation of the data set that will be used in the subsequent modelling phase of the process (Colleen, 2006).

Attribute reduction and attribute's variable reduction are used in the study. Quality feature (attribute) is removed from the data matrix by the SAV genetic reducer algorithm (genetic algorithm) with Rosetta which is a toolkit for analyzing tabular data within the framework of rough set theory. In addition, it is decided to remove the shift attribute, department attribute, and fault type attribute from the data matrix after checking the basic statistical solutions. The shift, fault type and department of the process are not effective in employee performance. Thus, the number of attributes is reduced from 12 to 8. Attribute relevance analysis is used with the information gain technique reduce the number of qualitative variables. Table 1 summarizes the results of the attribute relevance analysis. The accepted variables' gain ratios and information gains are higher than the averages. It is also shown in Table 1 that there is an important decrease in the numbers of variables after reduction.

Table 1: Attribute relevance results

Attribute Name	Average Gain Ratio	Number of Variables	Number of Accepted Variables
Fault reason	9,8105026199E-02	104	12
Design	3,7253001107E-02	200	43
Color	0,25233221877	18	8
Department	0,22613610383	10	5
Workbench	0,33821371848	10	6
Employee	0,1436797671	36	18

Anomaly detection model (Ciflikli C. and Kahya-Özyirmidokuz E, 2010) is used to identify outliers, or unusual cases in the data. Two groups are detected. The records which are greater than the average anomaly index level (1.87662) are selected as anomaly records. In this way, 514 of the records are eliminated with the anomaly detection algorithm.

Correlation and regression analyses are used to understand if a relationship occurs. The categorization of the non-binarized data matrix is established. Web graphs are formed. An inverse relation of 0.850 is found between the design attribute and the employee attribute. Figure 3 shows the web graph of the design and the employee. The regression analysis is established and the relation is shown in Equation (1).

$$Employee = 16.9 + 0.0032\ Design \quad (1)$$

It was decided not to use employee and design attributes together in the DM model because of the high correlation between them.

4. Decision tree induction

After preprocessing, the data matrix is reduced to 3,917 records and the attributes are shown in Table 2. Modeling is performed with C4.5(Quinlan, 1993), which is one of the popular DT modelling algorithms, an extension of an earlier well known algorithm, ID3. The basic algorithm for DT induction is a greedy algorithm that constructs DTs in a top-down recursive divide and conquer manner (Han and Kamber, 2001).

ID3 utilizes entropy criteria for splitting nodes. Given a node *t*, the splitting criterion used is $Entropy(t) = \sum_i -p_i \log p_i$ where p_i is the probability of class *i* within node *t*. An attribute and split are selected to minimize entropy. Splitting a node produces two or more direct descendants. Each child has a measure of entropy. The sum of each child's entropy is weighted by its percentage of the parent's cases in computing the final weighted entropy used to decide the best split. In C4.5, given node *t*, the splitting criterion used is the *GainRatio(t)=gain(t)/SplitInformation(t).* This ratio expresses the proportion of information generated by a split that is helpful for developing the classification, and may be thought of as a normalized information gain or entropy measure for the test. A test is selected that maximizes this ratio, as long as the numerator (the information gain) is larger than the average gain across all tests. The numerator in this ratio is the standard information entropy

difference achieved at node *t*, expressed as $gain(t) = \inf o(T) - \inf o_t(T)$, where $\inf o(T) = -\sum_{i=1}^{k} C_i / C_T$, and $\inf o_t(T) = \sum_{i=1}^{s} (T_i / T) \times \inf o(T_i)$ (Apte and Weiss, 1997). The choice of DM tool must be based on the application domain and its supported features (Hui, Jha, 2000).

Figure 2: Web graph of design and employee

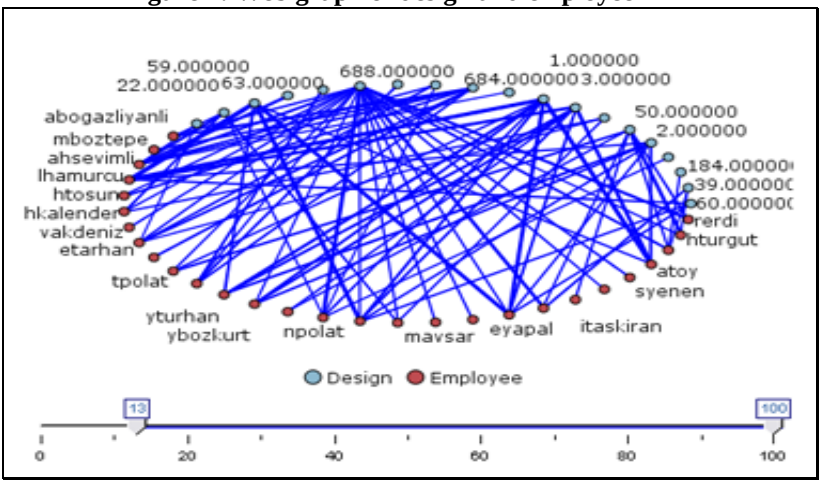


Table 2: Attribute variables' frequency

Before Preprocessing		**After Preprocessing**	
Attribute's Name	**Number of Initial Variables**	**Attribute's Name**	**Number of Final Variables**
Fault reason	104	Fault reason	12
Design	200	Design	43
Color	18	Color	8
Workbench	10	Workbench	6
Month	Discrete	Month	12
Size	13	Size	13
Employee	36	Employee	18
Shift	3	-	-
Fault Type	8	-	-
Quality	200	-	-
Department	9	-	-

The data matrix includes 6 inputs and a target attribute, which is the employee attribute. The presented DT model is acceptable with 7 tree depths. The relationships of a large number of input variables to target the attribute's variables are found. Moreover, 460 if-then rules are generated to express the process. The following examples illustrate one of the rules belonging to the related employee:

- Rule for Employee1
  if workbench = 1 and Month in [1.000 2.000 3.000 4.000 5.000 6.000]
  and employee fault reason in ["Reason5"] and Design in [3.000]
  and size in ["100X200"] then employee1
- Rule 1 for Employee14
  if workbench=6 and Month in [1.000 2.000 3.000 4.000 5.000 6.000 7.000 8.000] and employee fault reason in ["Reason1"] then Employee14

The data are partitioned as training and test data which from 20% of the data. Ten fold cross-validation accuracy evaluation is used to train and test the data matrix. The accuracy ratio of the model is 74.86%.

5. Conclusions

DM is one of the emerging technologies in the field of data analysis tools. It has a significant impact in HR in discovering patterns and information hidden in databases in order to use this information to gain knowledge to help decision makers to develop strategies and future plans for the labor market (Alsultanny, 2013).

We discovered useful patterns from the fabric data to highlight employee efficiency. For this reason, the DT method was used. DT provides an effective approach to identify the employee performance. Rules were generated. The efficiency of employees and the reasons for this situation were detected with the model. The faults of the employees were discovered. The acquired tree model is different from existing employee efficiency detection methods that have problems such as being difficulty to understand. This study presents a logical model that can be easily understood.

6. References

- Aiolli, F.; de Filippo, M.; Sperduti, A. (2009) Application of the preference learning model to a HR selection task. *I*n: *IEEE*

symposium on computational intelligence and DM (CIDM) 2009: Nashville, TN, p. 203–210.

- Alsultanny, Y. A. (2013) Labor market forecasting by using DM. *Procedia Computer Science,* 18, p.1700 – 1709.
- Apte, C.; Weiss, S. (1997) DM with DTs and decision rules. *Future Generation Computer Systems,* 13, p.197-210.
- Aviad, B.; Roy, G. (2011). Classification by clustering DT-like classifier based on adjusted clusters. *Expert Systems with Applications,* 38(7), p. 8220–8228.
- Ciflikli, C.; Kahya Ozyirmidokuz, E. (2008) A DM application in carpet manufacturing industry to determine employee inefficiency. *6th international symposium on Intelligent manufacturing systems* (IMS) 2008: Sakarya, Turkey, p.14.
- Ciflikli, C.; Kahya Ozyirmidokuz, E. (2010) Implementing A DM solution for enhancing carpet manufacturing productivity. *Knowledge Based Systems,* 23, p.783-788.
- Ciflikli, C.; Kahya Ozyirmidokuz, E. (2012). Enhancing product quality of a process. *Industrial Management & Data Systems*, 112 (8), p.1181-1200.
- Han, J.; Kamber, M. (2001) *DM: Concepts and Techniques*, Morgan Kaufmann Publishers: USA.
- Harding, J. A.; Shahbaz, M.; Srinivas; Kusiak, A. (2006) DM in manufacturing: A review, Journal of Manufacturing Science and Engineering, *Manufacturing Engineering Division of Asme*, 128, p.969-976.
- Hui, S. C.; Jha, G. (2000) DM for customer service support. *Information & Management,* 38, p.1-13.
- Juan, L. (2009) Early warning model research of state-owned enterprises' human capital risks based on improved neural network. In: *International Conference on Future BioMedical Information Engineering*, Sanya, China, p.197–201.
- Karahoca, A; Karahoca, D.; Kaya, O. (2008) DM to cluster human performance by using online self regulating clustering method. In: Proceedings of the 1st WSEAS international conference on multivariate analysis and its application in science and engineering (MAASE) 2008: İstanbul Turkey, p. 198–203.

- Larose, D. T. (2005) Discovering Knowledge in Data: An Introduction to DM, Wiley: USA.
- Lee, S. J.; Siau, K. (2001) A review of DM techniques. *Industrial Management and Data Systems,* 101(1), p.41-46.
- Li, W.; Xu, S.; Meng, W. (2009) A risk prediction model of construction enterprise HR based on support vector machine. In: *Second international conference on intelligent computation technology and automation* (ICICTA) 2009: Human, China, p. 945–948.
- Lockamy, A.; Service, R. W. (2011) Modeling managerial promotion decisions using Bayesian networks: An exploratory study. *Journal of Management Development*, 30, p.381–401.
- Ramesh, B. (2001) GPR: A DM tool using genetic programming. *Communications of the Association for Information Systems* 5: 1–36.
- Sivaram, N.; Ramar, K. (2010) Applicability of clustering and classification algorithms for recruitment DM. *International Journal of Computer Applications*, 4(5), p.23–28.
- Soares, C.; Ghani, R. (2010) DM for Business Applications: Introduction. In: Soares C, Ghani R (eds). *DM for Business Applications* 2010. IOS Press: Netherlands, p. 1-34.
- Strohmeier, S.; Piazza, F. (2013) Domain driven DM in HR management: A review of current research. *Expert Systems with Applications*, 40, p.2410–2420.
- Thissen-Roe, A. (2005) Adaptive selection of personality items to inform a neural network predicting job performance. phD thesis, University of Washington.
- Wang, K. (2007) Applying DM to manufacturing: the nature and implications. *Journal of Intelligent Manufacturing*. 18: 487–495.
- Zhang, D.; Deng, J. (2011) The DM of the HR data warehouse in university based on association rule. *Journal of Computers*, 6, p.139–146.
- Zhao, X. (2008) An empirical study of DM in performance evaluation of HR management. In: *International symposium on intelligent information technology application workshops (IITAW) 2008*, Shanghai, p. 82–85.