Employee Turnover: A Novel Prediction Solution with Effective Feature Selection

HSIN-YUN CHANG
Department of Business Administration
Chin-Min Institute of Technology
110 Hsueh-Fu Road, Tou-Fen, Miao-Li 305
TAIWAN, R.O.C.
ran_hsin@ms.chinmin.edu.tw

Abstract: - This study proposed to address a new method that could select subsets more efficiently. In addition, the reasons why employers voluntarily turnover were also investigated in order to increase the classification accuracy and to help managers to prevent employers' turnover. The mixed feature subset selection used in this study combined Taguchi method and Nearest Neighbor Classification Rules to select feature subset and analyze the factors to find the best predictor of employer turnover. All the samples used in this study were from industry A, in which the employers left their job during 1st of February, 2001 to 31st of December, 2007, compared with those incumbents. The results showed that through the mixed feature subset selection method, total 18 factors were found that are important to the employers. In addition, the accuracy of correct selection was 87.85% which was higher than before using this feature subset selection method (80.93%). The new feature subset selection method addressed in this study does not only provide industries to understand the reasons of employers' turnover, but also could be a long-term classification prediction for industries.

Key-Words: - Voluntary Turnover; Feature Subset Selection; Taguchi Methods; Nearest Neighbor Classification Rules; Training pattern

1 Introduction

Human resource is the most important asset for a company to be competitive. Thanks to liberalization on the labor market, it becomes possible for an employee to leave his job and when employees of a department change frequently, the employee allocation will be more efficient. However, having excess employees leave their jobs will influence the morale of the companies, called Snowball Effect [35] (a man's leaving induces his colleagues to leave one by one). This has a great effect upon a company's operation. Innovation of a product could be duplicated, but good teamwork and employees cannot be duplicated. The loss of good employees can diminish a company's competitive advantage and furthermore lead to a reduction in output and quality.

To avoid a huge loss and wide influence, a company has no alternative but to decrease or slow its employee's turnover rate and to find out the true causes for employees' turnover.

As a result, to assist companies in building an early warning system of predicting their employees' leaving, the investigator attempted to find out the causes through a hybrid feature selection model.

From 1970's, feature selection became an important subject in academic fields, such as data mining, business intelligence [8], machine learning, and pattern recognition [23,27,31,41]. At the same time, it was applied extensively to every question fields, including intrusion detection [2], genomic analysis [4], customer relationship management [5], image retrieval [3], and text categorization [6].

Feature selection, also known as feature subset selection [7,43] is one of the most common methods applied to data preprocessing. With the boom and burst in information technology, including network technology, database technology and so on, the large amount of all collected data has come to an extent where people are not capable of dealing with it.

Due to the explosion of information, many methods in machine learning and data mining have been introduced in order to pick out relevant information or knowledge.

From the viewpoint of academic research or practical application, data preprocessing [9] is one of the keys to the success of machine learning and data mining. Of all data preprocessing methods, feature selection is the most popular and important method.

In many fields of classifying question, feature selection is taken seriously. Therefore, the investigator attempted to propose a new feature selection method with better identification. This method can simplify calculating process and time in classifiers and classification method, can help to understand the relation between cause and effect of classification questions of staff turnover, and can create benefits when companies perform future human resource strategy and arrange their organizations.

2 Turnover and Turnover Intention

In order to understand the causes for employees' leaving in manufacturing industry, it was necessary to have a clear definition of turnover.

2.1 Turnover

In Emery and Trist's research, Turnover has already been defined as follows. When an individual entered a company, the interaction between the company and the individual was supposed to increase. If the interaction could not increase to an appropriate extent, the individual's past experience would turn to be so-called Guiding Crisis and the individual would leave eventually [10].

Bluedorn gave a new definition of turnover in 1982. Turnover or turnover process did not only mean an individual left the company. It meant the individual stopped playing a role in the company and left the relevant areas of the company [11,12].

Recently, as the quick development of high-tech industries, such as e-information, e-communication, output value, numbers of employee and research funding increased gradually and it seems become the major power of economic growth in Taiwan and also the major output industry. Therefore, rise and decline of high-tech industries influences the competition and economical growth of industries in Taiwan. In addition, it also affects whether or not the economic development in Taiwan would grow better. Because high-tech industries is an industry with knowledge, the key point is the quality of employees and if there is enough supply of employees. However, employees who work in high-tech industries usually do not want to load much long-term pressure and due to some other factors, their leaving has a large influence in the industries.

In this study, employees who work in conventional industries were used as samples. First aim was to construct a model of turnover, and then to predict in what circumstances, employees would choose to quit through the proposed method. Further, the results from this study would be a great contribution of human resource and organization of industries in the future.

In order to understand the reasons of why industries employees leave their job, first of all, the definition of 'turnover' should be clearly defined. Therefore, the meaning of 'the meaning of turnover', 'types of employees' turnover' and 'effect of employees' turnover' were described as the meaning of job leaving is turnover. It includes that transfer their job from one location to another (special transfer), or from one job to another (industrial transfer), or from one industry to another (industrial transfer). It could also happen from one organization to another (e.g., industry and organization), input and output of employee. Narrative definition of employee turnover means that people quit their current job. Price [3] defined 'employee turnover' as that personal move over the boundary of the other organization, and this could mean entering to or leaving an organization. Ferratt and Short [32] defined it as break off relationship between employees and, no matter who cause this it could be called employee turnover.

Traditionally, there are two types of turnover [33]:
(1) Voluntary turnover:

Most of job movement is addressed by employees, a personal choice of employee turnover.

(2) Involuntary turnover

The employee turnover is not controllable, such as retire, dismiss or death. Under some circumstances, although labor addresses it, it still should be taken into account as involuntary employee turnover, because the employees have no choice.

From two types of employee turnover mentioned above, voluntary leaving is the most important issue that industries should think about. King and Seith [34] suggested that it is impossible for an organization to avoid employee turnover, for instance, marriage, and pregnancy. Thus, whether organization could avoid employee turnover was investigated in this study.

Through the study of 'voluntary leaving', we could understand the problem of management in industries and also normal condition in organizations. For example, while the rate of employee turnover is high, several factors should be concerned, such as manager, boring work, lack of achievement, bad work environment. Thus, there are varieties of factors that could cause turnover.

To sum up scholars' viewpoints, turnover means an employee leaves the company completely. It also means the relations between labor and capital breaks off [13]. No matter employees leave voluntarily or

involuntarily, if, in the entire process, interaction cannot increase to an expected extent or a negative outcome to the job happens, it can also mean turnover.

2.2 Turnover Intention

Turnover intention (TOI) is the best factor for predicting turnover [14-20]. Turnover intention means the strength of intention an individual has to leave his present job and look for another job opportunity [21]. Many studies show that employee turnover intention has strong relation to organizations [22]. Employees have an intention or plan to leave their jobs because of the work they are doing and the organization they are in.

Accordingly, turnover intention is a significant factor in predicting turnover. By examining relevant research, this research is going to offer an overall understanding of employees' turnover intention and further predict the key factors that influence employees' turnover.

3 Feature Selection Model

3.1 Definition

Turnover intention (TOI) is the best factor for predicting turnover [14-20]. Turnover intention means the strength of intention an individual has to leave his present job and look for another job opportunity [21]. Many studies show that employee turnover intention has strong relation to organizations [22]. Employees have an intention or plan to leave their jobs because of the work they are doing and the organization they are in.

Accordingly, turnover intention is a significant factor in predicting turnover. By examining relevant research, this research is going to offer an overall understanding of employees' turnover intention and further predict the key factors that influence employees' turnover.

Prior to feature selection, a set of training instances or patterns (training set) is given, where each instance is represented by n features and an output label. Many pattern classification have been investegat to classify new, unseen instances based on extracting useful knowledge from the training set. Theoretically, all features of each instance will be considered for the purpose of classification. But real-world classification tasks usually contain irrelevant or redundant features, which may degrade the classification accuracy. Consequently, many feature subset selection approaches [3,4,7] have been developed to reduce the dimensionality in pattern classification. In other words, irrelevant or redundant features will be removed from the original feature set .

Feature subset selection is a process that selects important or relevant features from the original feature set. It is also a search problem [26], where each search state in the search space identifies a possible feature subset. Feature subset selection offers many advantages for pattern classification. Firstly, the cost of gathering training or unseen instances can be reduced. Secondly, pattern classification models can be constructed faster. Furthermore, classification accuracy and the comprehensibility of the learning models can be improved. If each instance contains n features, the search space will be composed of 2n candidate feature subsets. Obviously, exhaustive search through the entire search space has a very high computational cost and thus is usually unfeasible in practice, even for medium-sized n [24]. Consequently, it is difficult to select a best feature subset for pattern classification from the entire search space feature selection is a process of picking out a particular feature subset from feature sets [7, 26]. In order to make sure the feature Subset is optimal, a specific subset evaluation is necessary.

3.2 Purposes of Feature Selection

Feature subset selection is generally carried out in four steps [24,25,26]. (1)The search starting point in the search space; (2)A generation rule with search strategies to generate the next candidate feature subset; (3)An evaluation function to evaluate each generated feature subset; (4)A stopping criterion to determine when to halt the selection process.

As a determinative and principal step for feature subset selection, the search starting point in the search space is used to decide the direction of the search [26]. Generally, the feature subset search procedure can start with no features (for example, sequential forward selection method [30]) or all features (for example, sequential backward elimination method [30]). Accordingly, features are successively added or eliminated (i.e., deterministic heuristic search). In these two cases, sub-optimal feature subsets are often obtained because of successive additions or eliminations of features. In another approach, random sampling [36], the feature subset search procedure can start with a random subset of features. This method can help the search procedure to escape from local maximums [26] However, inconsistent final feature subsets may be derived from different runs [26]. In other words search starting point determination plays a vital role

here and will significantly affect the performance of the corresponding feature subset selection method.

4 A Novel Approach to Hybrid Feature Selection Model

4.1 Taguchi Methods

As a well-known robust experimental design approach, the Taguchi method [37,38] uses two principal tools, the orthogonal array and the signal-to-noise ratio (SNR), for the purpose of evaluation and improvement. Consider that a specific object domain contains q design parameters (or factors). Orthogonal arrays are primarily used to reduce the experimental efforts regarding these q different design factors. An orthogonal array can be viewed as a fractional factorial matrix that provides a systematic and balanced comparison of different levels of each design factor and interactions among all design factors. In this two-dimensional matrix, each column specifies a particular design factor and each row represents a trial with a specific combination of different levels regarding all design factors. In the proposed method, the well-known two-level orthogonal array is adopted for feature subset selection.

For example, an orthogonal array can be created for a specific object domain that contains 15 design factors with two levels (i.e., level 1 and level 2). Notably, by using the two-level orthogonal array, only 16 experimental trials are needed for the purpose of evaluation and improvement. By contrast, all possible combinations of 15 design factors (i.e., 215) should be taken into consideration in the full factorial experimental design, which is obviously often inapplicable in practice.

Once the orthogonal array is generated, the observation or the objective function of each experimental trial can be determined. Accordingly, the signal-to-noise ratio (SNR) is used to evaluate and optimize the design parameters (or factors) of the specific object domain. In general, two kinds of signal-to-noise ratios (SNRs), the smaller-the-better and the larger-the-better characteristics [37,38], are commonly considered for the evaluation task.

The signal-to-noise ratio (SNR) is used to measure the robustness of each design parameter (or factor). That is, "high quality" of a particular object domain can be achieved by considering each design parameter with a specific level having high signal-to-noise ratio (SNR).

The Taguchi method offers many advantages for robust experimental design. First, the number of experimental runs can be substantially reduced. Meanwhile, the significance of each design parameter regarding a particular object domain can be analyzed precisely. In the proposed method, the above two useful tools, the orthogonal array and the signal-to-noise ratio (SNR), are employed for feature subset selection.

4.2 Feature Selection Based on Taguchi Methods

In order to predict employees' turnover variables more sufficiently, this study proposed to find a hybrid model of feature selection.

Consider that a set of m labeled training instances V={v1, v2, …, vm} is given in a specific classification task. Each instance contains n features, which are represented as F = (f1, f2, …, fn). Using a pattern classification model with the leave-one-out cross-validation method [39,42], a classification accuracy can be determined or measured for the training set V and feature set F (denoted by ACC(V,F)). The leave-one-out cross-validation method assumes that each instance in V is used as the test instance once and other instances in V are used as the corresponding training instances. In other words, the pattern classification model [1] will be performed m times, with respect to m instances and n features in V. Accordingly, the average classification accuracy, denoted as ACC(V,F), can be obtained or measured and then is used in the proposed method.

Also, a corresponding classification accuracy ACC(V, Fi) can be determined (Fi=F-{fi}). The value of ACC(V, Fi) indicates the classification accuracy obtained by using the pattern classification model with the feature subset Fi. That is, feature fi is not considered to be in the feature space. In other words, the difference between ACC(V,F) and ACC(V, Fi), denoted as DIF(fi), points out the classification effectiveness of each feature fi in the pattern classification model. Here, feature fi is excluded from the original feature set F and the relationships or relevance among all other features are reflected. If the difference between ACC(V,F) and ACC(V, Fi) (i.e., DIF(fi)) is large enough, feature fi can then be viewed as a relevant and important feature. This is because the exclusion of feature fi from the original feature set F will significantly degrade the overall and baseline classification accuracy. Hence, feature fi should be considered first in feature subset selection. Alternatively, feature fi can be viewed as an irrelevant and meaningless feature if the difference

between ACC(V,F) and ACC(V, Fi)(i.e., DIF(fi)) is small.

The proposed feature selection method is on the basis of Taguchi Methods and matched with a sorting method that to examine the efficiency (accuracy of sorting) in order to decide the quality of each sub feature and improve the accuracy.

The author of this study suggested a hybrid model which on the bases of Taguchi Methods. It could be hybrid model [4], in other words, the authors tried to combine Filter Model [28,29] and Wrapper Model [30,24] to evaluate the feature subset.

Therefore, the efficiency of the feature subset evaluation was examined through pre-selected sorting method to decide the advantage or disadvantage of each feature subset. The strategies are as follows:

Restated, the author assumed that there are m training samples, representing as V= {v1, v2… vm}. Each sample had n categories, represented as F = (f1, f2, …, fn).

The detailed steps are described as follows:

Step 1. The author firstly assumed a two-level orthogonal table L with n categories. Each experiment j in the orthogonal table L has level 1 and 2 represented the category i which could be selected or not in feature subset Sj.

Step 2. The pattern classification model and Leave-one-out cross-validation were used for each feature subset, Sj to find their average accuracy, represented as ACC(V, Sj). ACC (V, Sj) could be the numbers that were seen in the experiment j in the orthogonal table L.

Step 3. According to the observed numbers in the orthogonal table L, the level 1 and 2 for each feature has a relative SNR value.

Step 4. For those features that can select higher SNR value of level 1 than of level 2, were represented as S, which was used as the last feature sub set.

The orthogonal table could be seen as a variable (matrix) that supplies a comparison which is systemic and symmetrical in order to explore the relationship between all factors. In other words, the two dimension matrix aims to reduce the time consuming and costs. Subsequently, the nearest neighborhood rule and signal-to-noise ratio were used as the evaluation criteria for feature selection.

4.3 Best evaluation of feature – nearest neighborhood rule and signal-to-noise ratio (SNR)

According to the nearest neighborhood rule and leave-one-out method could find the average accuracy of feature subset Sj for each experiment j, labeled as ACC (V, Sj). The leave-one-out method aims to see each sample as a tested sample and the others as the relative samples.

Therefore, the nearest neighborhood rule would be ran m times (sample size). Subsequently, the average accuracy should be calculated to evaluate the efficiency of feature subset Sj, namely, the level of each feature and the best SNR value are related to the efficiency.

Thus, the higher the better will be used to calculate SNR value. Because in the sorting samples, the higher accuracy the better, indicating that the feature is the suggested factor for feature subset. In the opposite, in feature i, if the SNR of level 2 is higher than of level 1, the feature will be suggested to filter out from the original feature set F.

5 Result analysis

To examine the accuracy of the proposed hybrid feature selection method, the investigator had a test on the prediction of turnover classification by using collected data of 881 employees in manufacturing field.

881 employees who work in manufactory industries were recruited through the hybrid model mentioned above. There were 44 features selected. All of these features were considered as feature variables and analyzed based on Taguchi Methods feature selection method.

5.1 Forecast verification of the Classification of employees' turnover

(I) 881 data were randomly selected (441 data were assigned as training samples and 440 data were test samples). Subsequently, 441 data were used as test sample, and 440 data were training samples. When the proposed method is not used, the classification accuracy of present employees is 89.03% (536/602*100%) and the classification accuracy of left employees 63.44% (177/279*100%), the overall accuracy is 80.93% (774/881*100%). By contrast, when the proposed method is used, the classification accuracy of present employees is 93.36% (562/602*100%) and the classification accuracy of

left employees 75.99% (212/279*100%), the overall accuracy is 87.85% (713/881*100%). As a result, the proposed method yields superior performance.

(II) Average results from ten experiments: 881 data were selected randomly (training sample 50%, testing sample 50%). The Taguchi Methods method was used for feature selection. Each time, the author calculated the average accuracy, and also estimated the average accuracy of ten times. For these classification problems, the classification abilities or accuracies of all search starting points obtained by random sampling and by using the proposed method are79.93% and 87.39%.
The results showed that the new feature selection model suggested in this study could improve the industry to predict whether an employee has tendency to turnover. This could simplify the sorting procedure and decline the costs. In addition, the method suggested in this study could help for the causal relationship. Furthermore, this model could be used as a longitudinal method for industry.

1.gender
2.under 22 y/o
3.education level-college
4.marriage
5.resident
6.full- or part time
7.Salary under 25000
8.department
9.position
10. pre-experience
11.seniority
12.average validated credit
13.average age of their children
14.whether the partner is working or not
15.pursuing further education?
16.over working hours
17.involved in the activity held by the company
18.sick leave
19.seniority is less than 1 year
20.average validated credit

Table 5.1 Features related to turnover categories

5.2 Factor analysis of the predict variables of employees' turnover

Table 5.1 shows those selected variables which were selected by Taguchi Methods hybrid model. This was designed to analyze training program on the basis of Taguchi Methods model, and additionally, select those important and related variables which are relevant to employees' turnover.

5.3 Additional Experimental Results

To demonstrate the performance of the proposed method, various real datasets [40] were used for performance comparison. Table 5.2 represents the main characteristics of the datasets.

Let X denote the search starting point for feature subset selection obtained by using the proposed method and Y denote the best feature subset in a specific classification task. A match ratio (MR) regarding X and Y can be introduced as follows:

$$match\ ratio\ (MR) = \frac{|X \cap Y|}{|X|} \qquad (1)$$

where and denote the number of features in feature sets X and Y, respectively.

Obviously, a search starting point X with high match ratio MR means that nearly all features in X (i.e., nearly features in the original feature set) are 'also' included in the best feature subset Y. By using the corresponding search starting point X in the search procedure for feature subset selection, the final best feature subset Y can then be approached more effectively.

Table 5.3 represents the match ratios MRs of the search starting points obtained by random sampling [36] and by using the proposed method, with respect to the above-mentioned classification tasks or datasets. For these classification domains, the average match ratios (MRs) of the search starting points obtained by random sampling and by using the proposed method are 61.52% and 94.55%, respectively. In other words, as expected, nearly all features (94.55%) in the search starting point obtained by using the proposed method are also included in the final best feature subset of a specific classification task.

Classification task	Number of instances	Number of classes	Number of features and their types
Car	1728	4	6 (6-S)
Corral	16	2	6 (6-S)
Echocardiogram	74	2	10 (2-S, 8C)
Echoi	336	8	7 (7-C)
Glass	214	6	9 (9-C)
Haberman	306	2	3 (3-C)
Hayesroth	132	3	4 (4-S)
Hcleveland	303	5	12 (8-S, 5-C)
Iris	150	3	4 (4-C)
Liver Disorder	345	2	6 (6-C)
Lymphography	148	4	18 (15-S, 3-C)
Monk1	432	2	6 (6-S)
Monk2	432	2	6 (6-S)
Monk3	432	2	6 (6-S)
Nursery	12960	5	8 (8-S)
Pageblock	5473	5	10 (10-C)
Pimadiabetes	768	2	8 (8-C)
Postoperative	90	3	18 (18-S)
Primarytumor	339	21	17 (17-S)
Segment	2310	7	18 (18-C)
Shuttle	43500	7	9 (9-C)
Solarflare	323	6	12 (12-S)
Tae	151	3	5 (4-S, 1-C)
Tictactoe	958	2	9 (9-S)
Voting	435	2	16 (16-S)

C: Continuous, S: Symbolic

Table 5.2.
Details of experimental classification tasks

Classification task	The match ratio MR of the search starting point obtained by random sampling	The match ratio MR of the search starting point obtained by using the proposed method
Car	66.7%	100.0%
Corral	33.3%	100.0%
Echocardiogram	40.0%	100.0%
Echoi	100.0%	100.0%
Glass	60.0%	100.0%
Haberman	50.0%	100.0%
Hayesroth	50.0%	100.0%
Hcleveland	42.9%	85.7%
Lenses	50.0%	100.0%
Liver Disorder	66.7%	100.0%
Lymphography	44.4%	77.8%
Monk1	66.7%	100.0%
Monk2	100.0%	100.0%
Monk3	33.3%	100.0%
Nursey	50.0%	100.0%
Pageblock	80.0%	80.0%
Pimadiabetes	75.0%	100.0%
Postoperative	55.6%	87.5%
Primarytumor	66.7%	100.0%
Segment	66.7%	77.8%
Shuttle	60.0%	80.0%
Solarflare	83.3%	100.0%
Tae	66.7%	100.0%
Tictactoe	80.0%	100.0%
Voting	50.0%	75.0%
Average	61.52%	94.55%

Table 5.3.
The match ratios MRs of the search starting points obtained by random sampling and by using the proposed feature selection method

6 Conclusions

The hybrid model suggested in this study was the combination with Taguchi Methods and the nearest neighborhood rule. Features with higher classification effectiveness are more important and relevant for the specific classification task. The set of these important and relevant features is thus considered as the search starting point and is expected to have high relevance to the best feature subset for feature subset selection. The search starting point for feature subset selection is determined by somewhere in the middle of the search space. Besides, in order to examine the efficiency of this model, the data base contained those employees who were present from 1st of February, 2007 to 31st of December, 2007 supplied by industry A were analyzed.

The results showed that the model used in this study could be the best model of categorizing, and the accuracy was 87.85%.

The results showed that the best model of turnover prediction was the Taguchi Methods combined with the nearest neighbor rule. This model could help individual industry to establish their database in order to investigate which factors could be used as prediction for employees' turnover. Because, there may be some signs before an employee really apply for turnover, the key point is that whether manager could notice or not. Consequently, this system is suggested for the industry to establish their own system to predict employees' turnover.

Furthermore, the author suggested that industries could analyze their employees' data through the feature selection method mentioned in this study regularly. The output could be as reference for the manager. In addition, the data base should be updated regularly. Moreover, the system of turnover prediction should be monitored and modified regularly in order to explore any new variable that changes by the environment.

In addition, in some classification problem domains, some additional features probably need to be included in the proposed search staring point or some features, such as redundant features, probably need to be excluded from the proposed search staring point to obtain the final best feature subset. To deal with such cases, the proposed feature ranking method with search starting point determination can be incorporated with other existing feature subset selection methods, such as sequential floating search methods, sequential forward or sequential backward selection methods. Based on the concepts of these existing sequential methods, features can be successively added or eliminated for feature subset selection.

References:

[1] M. Ben. Bassat, "Pattern Recognition and Reduction of Dimensionality," *Handbook of Statistics-II*, P.R. Krishnaiah and L.N. Kanal, eds.,1982, pp. 773-791, North Holland.

[2] W. Lee, S.J., Stolfo, and K.W. Mok, "Adaptive Intrusion Detection: A Data Mining Approach", *AI Rev*., vol. 14, no. 6, 2000, pp. 533-567.

[3] D.L. Swets, and J.J. Weng, "Efficient Content-Based Image Retrieval Using Automatic Feature Selection", *IEEE Int'l Symp. Computer Vision*, 1995, pp. 85-90.

[4] E. Xing, M. Jordan, and R. Karp, "Feature Selection for High-Dimensional Genomic Micro array Data", In *Proc. 15th Int'l Conf. Machine Learning*, 2001, pp. 601-608.

[5] Ng K.S. and H. Liu, "Customer Retention via Data Mining", *AI Rev*., vol. 14, no. 6, 2000, pp. 569-590.

[6] K. A. Nigam, K. S. Mccallum, Thrun, and T. Mitchell, "Text Classification from Labeled and Unlabeled Documents Using EM", *Machine Learning*, vol. 39, 2000, pp. 103-134.

[7] J. G. Dy, and C.E. Brodley, "Feature Subset Selection and Order Identification for Unsupervised Learning", in *Proc. 17th Int'l Conf. Machine Learning*, 2000, pp. 247-254.

[8] Carlo Dell' aquila, Francesco Di Tria, Ezio Lefons, Filippo Tangorra, "Business Intelligence Systems: A Comparative Analysis", *Wseas Transactions on Information Science and Applications*, Issue 5, Vol. 5, May 2008.

[9] F. W. Famili, M. Shen, R. Weber, and E. Simoudis, "Data Preprocessing and Intelligent Data Analysis", *Intell. Data Anal*. 1(1-4), 1997, pp. 3-23.

[10] F.E. Emery, and E.C. Trist, "The Causal Texture of organizations". *Human Relations*, vol. 18, 1965, pp. 21-31.

[11] J. L. Price, *The study of turnover*, Ames: Iowa State University Press, 1997.

[12] A. C. Bluedron, "The theories of turnover: Causes effects and eaning", *Research in the Sociology of Organization*, vol.35, 1982, pp. 135-153.

[13] G. H. Ferguson, and W.F. Ferguson, "Distinguishing Voluntary from Involuntary, Nurse Turnover", *Nursing management*, 17(12), 1986, pp. 43-44.

[14] J. E. Newman, "Predicting absenteeism and turnover: A field comparison of fishbein's model and traditional job attitude measure",

Journal of Applied Psychology, 59, 1974, pp. 610-15.

[15] W. H. Mobley, R. W. Griffeth, H. H. Hand, and Meglino, B. M. "Review and Conceptual Analysis of the Employee Turnover Process", *Psychological Bulletin*, vol. 86, no. 3, 1979, pp. 493-522.

[16] C. E. Michaels, and P.E. Spector, "Causes of employee turnover：A test of the mobley, griffeth, hand and meglino model" , *Journal of Applied Psychology*, vol. 67, no.1, 1982, pp. 53-59.

[17] R. P. Steel, and N. K. Ovalle, "A Review and Meta-analysis of Research on the Relationship Between Behavioural Intention and Employee Turnover", *Journal of Applied Psychology* (69), 1984, pp. 673-686.

[18] J. M. Carsten, and P. E. Spector, "Unemployment, job satisfaction, and employee turnover: a meta-analytic test of the Muchinsky model", *Journal of Applied Psychology*, 72, 1987, pp. 374-381.

[19] R. P. Tett, and J. P. Meyer, "Job Satisfaction, Organization Commitment, Turnover Intention, and Turnover: Path Analyses Based on Meta-Analytic Findings", *Personal Psychology* (40), 1993, pp. 259-291.

[20] D. S. Carlson, K. M. Kacmar, and L. P. Stepina, "An Examination of Two Aspects of Work-Family Conflict: Time and Identity", *Women in Management Review*, 10(2), 1995, pp. 17-25.

[21] R. D. Caplan, and K.W. Jones, "Effects of work load, role ambiguity and personality Type A on anxiety, depression, and heart rate", *Journal of Applied Psychology*, 60, 1975, pp.713-719.

[22] W. H. Mobley, "Intermediate Linkages in the Relationship Between Job Satisfaction and Employee Turnover", *Journal of Applied Psychology*, vol. 62, 1977, pp. 237-240.

[23] Salama Brook and Zaher Al Aghbari, "Classification of Personal Arabic Handwritten Documents", *Wseas Transactions on Information Science and Applications*, Issue 6, Vol. 5, June 2008.

[24] R. Kohavi, and G. H. John, "Wrappers for feature subset selection". *Artificial Intelligence*, 97, 1997, pp. 273-324.

[25] A. Blum, and P. Langley, "Selection of relevant features and examples in machine learning". *Artificial Intelligence*, 97, 1997, pp. 245-271.

[26] H. Liu, and L. Yu, "Toward Integrating Feature Selection Algorithms for Classification and Clustering". *IEEE Trans. Knowl. Data Eng*. 17, 2005, pp. 491-502.

[27] Ran Jin and Zhuojun Dong, Face Recognition based on Multi-scale Singular Value Features, *Wseas Transactions on Computers*, Issue 1, Vol. 8, January 2009.

[28] M. A. Hall, "Correlation-Based Feature Selection for Discrete and Numeric Class Machine Learning", in *Proc. 17th Int'l Conf. Machine Learning*, 2000, pp. 359-366.

[29] L. Yu, and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", In *Proc. 20th Int'l Conf. Machine Learning*, 2003, pp. 856-863.

[30] H. Liu , and Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*. Boston: Kluwer Academic,1998.

[31] R. O. Duda, and P. E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, 1973.

[32] T. W. Ferratt, and L. E. Short, "Are Information Systems People Different: An Investigation of Motivational Differences", *MIS Quarterly*, 1986, pp. 377-387.

[33] M. Igbaria, "Career Orientations of MIS: An Empirical Analysis", *MIS Quarterly*, June, 1991, pp. 151-169.

[34] R. King, and V. Seith, "The Impact of Socialization on the Role Adjustment of Information Systems Professionals", *Journal of Management Information Systems*, 14 (4), 1998, pp. 195-217.

[35] D. Krackhardt and L. W. Porter, "The snowball effect: Turnover embedded in communication networks," *Journal of Applied Psychology*, 71(1), 1986, pp. 50-55.

[36] G. Brassard, and P. Bratley, *Fundamentals of Algorithms*. New Jersey: Prentice Hall, 1996.

[37] G. Taguchi, *Introduction to Quality Engineering*. Tokyo: Asian Productivity Organization, 1986.

[38] Y. Wu, A. Wu, and G. Taguchi, *Taguchi Methods for Robust Design*. New York: ASME, 2000, pp. 7-10.

[39] M. Stone, "Cross-validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society B*, Vol. 36, 1974, pp. 111-147.

[40] C. L. Blake, and C. J. Merz, *UCI Repository of Machine Learning Databases* [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1998.

[41] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford, 1995.
[42] G. C. Cawley, and N. L. C. Talbot, "Efficient Leave-one-out Cross-validation of Kernel Fisher Discriminant Classifiers," *Pattern Recognition*, Vol. 36(11), 2003, pp. 2585-2592.
[43] J. Doak, "An Evaluation of Feature Selection Methods and Their Application to Computer Security," *Technical Report*, Univ. of California at Davis, Dept. Computer Science, 1992.