Expert Systems with Applications 39 (2012) 8844–8851

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals

Chin-Yuan Fan [a,*], Pei-Shu Fan [b], Te-Yi Chan [a], Shu-Hao Chang [a]

[a] Science and Technology Policy Research and Information Center, National Applied Research Laboratories, Taipei, Taiwan
[b] Department of Industrial Engineering and Management, China University of Science and Technology, Taipei, Taiwan

ARTICLE INFO

Keywords:
Turnover trend
Clustering analysis
Self-organizing map
Neural network clustering

ABSTRACT

This study applies clustering analysis for data mining and machine learning to predict trends in technology professional turnover rates, including the hybrid artificial neural network and clustering analysis known as the self-organizing map (SOM). This hybrid clustering method was used to study the individual characteristics of turnover trend clusters. Using a transaction questionnaire, we studied the period of peak turnover, which occurs after the Chinese New Year, for individuals divided into various age groups. The turnover trend of technology professionals was examined in well-known Taiwanese companies. The results indicate that the high outstanding turnover trend circle was primarily caused by a lack of inner fidelity identification, leadership and management. Based on cross-verification, the clustering accuracy rate was 92.7%. This study addressed problems related to the rapid loss of key human resources and should help organizations learn how to enhance competitiveness and efficiency.

1. Introduction

Improving the organization efficiency of human resources has been a key research topic for a long period of time. The purposes of this study were to discuss why Taiwan technology enterprises cannot retain talented employees and to find ways to prevent turnover of these employees and increase the competitiveness of the companies. Therefore, we use a discrete technique to discuss enterprise evolution. This study focused on employees between 20 and 39 years old who contribute to the high turnover in Taiwan. We examined turnover using a measurement scale based on real, local data. Reliability and validity tests were used to ensure data dependability. Other results can be revealed by applying computational intelligence clustering analysis techniques for data mining (i.e., self-organizing map (SOM) combined with back-propagation neural network (BPN)). SOM–BPN, which can be applied to reveal characteristics related to turnover trend clusters, is expected to offer reliable data to support the decision making of company policymakers.

In an effort to determine the clustering accuracy hit rate in SOM–BPN model, the variables were integrated and selected. This paper contributes to the forecasting literature by developing valid and reliable variables based on information obtained from prior literature and experts in the field. Moreover, this study developed an approach based on clustering analysis, SOM and neural network clustering, to determine the accuracy hit rate.

2. Problem statement and definitions

Past scholars have divided the turnover trend into "voluntary turnover", which may be individual or collective turnover, and "involuntary turnover", which may be due to retirement, death, misemployment, or a merger.

Dalton, Todor, and Krackhardt (1982) demarcated turnover in terms of the functions of an organization. Voluntary turnover was further divided into functional and non-functional turnover (Table 1). Most studies of organizations have paid more attention to non-functional turnover. In the present study, we emphasized voluntary turnover. Uncontrolled variables and poorly operating enterprises were not examined in the present study.

A high turnover rate or a large number of requests for resignation by exceptional employees will greatly influence the operation of an organization. Newman (1974), Kraut (1975), Mobley (1977), Mobley, Horner, and Hollingsworth (1978), Mobley, Griffeth, Hand, and Meglino (1979), Miller (1979) and Michaels and Spector (1982) all considered the best forecasting value for turnover to be the so-called "turnover trend". In addition, Porter and Steers (1973) indicated that the turnover trend is a potential phenomenon behavior whenever an employee experiences an unsatisfactory circumstance. Mobley (1977) showed that personal turnover trends are a determinant of retreating behavior. Nevertheless, the research of Mobley et al. (1978) indicated that the turnover trend

* Corresponding author.
E-mail addresses: cyfan@stpi.narl.org.tw, cyfan@mail.stpi.narl.org.tw (C.-Y. Fan).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.005

Table 1
The classification of turnover.

Organizational evaluation for employees		
	Excellent	Rotten
Personal evaluation for an organization		
Involuntary turnover	Staying in an organization	Employees are misemployed
Voluntary turnover	Employees leave organization (non-functional turnover)	Employees leave organization (functional turnover)

is the summation of unsatisfied work, turnover thought, employees looking for suitable jobs and the availability of potential jobs, which suggested that the forecasting target for turnover behavior can be derived through the turnover trend.

The present study focused on the period of high turnover in Taiwan from 2009. The current study was influenced by the variables and explanations of turnover provided by other studies in the literature within the last five years.

2.1. Data mining and clustering analysis method

Prior analyses of the turnover trend or other relative research topics mostly applied descriptive statistics combined with correlation analysis, variables analysis, analysis of variance (ANOVA) integrated with regression analysis, multiple hierarchical regression analysis and structural equation modeling analysis. Samuel and Chipunza (2009), Hao, Jung and Yenhui (2009), some scholars also applied back-propagation networks (BPNs), logistic regression and market basket data analysis to predict turnover trends, but these types of analyses have various defects. Back-propagation networks focus on the internal database of an enterprise for sources of data from registered employee turnover forms, which contain the opinions of the employee. The majority are not willing to provide the actual reasons for turnover, which makes the validity of BPN research questionable. In addition, the predictive ability of logistic regression cannot be validated. Although the concept and research results of market basket data analysis are acceptable, there are no similar objects that can be consulted; therefore, market basket data analysis cannot really help in predictions of turnover rate.

Because a suitable method of analysis did not exist, the present study applied a clustering analysis data mining technique. The present study utilized RMSE, which was applied through a SOM, to combine a two-phase clustering based on the BPN classification technique in clustering analysis.

3. Methodology

This study initially verified the questionnaire data sets. Questionnaire was based on the definition of each phase or the developed measurement tools used to establish question contents. Questionnaires were tested by a cross-section approach, and the procedure is shown in Fig. 1.

3.1. Data preprocessing

This section considers two parts of the study: data collection and model development. The second part involves investigating and selecting a suitable questionnaire. We attempted two types of systemic processes to transform multiform original data to effective and useful research data.

3.1.1. Define the research model according to the turnover trend

In this part, we assumed that there were five parts action associated with turnover trend. A cause-and-effect chart for each department is shown in Fig. 2. In this figure, work stress and organizational politics affect work satisfaction, and work satisfaction and leader promise directly affect organizational behavior. In addition, organizational behavior affects the turnover trend. The detailed flow chart is shown in Fig. 2.

3.1.2. Survey and filtering of valid questionnaires

The questionnaire survey for this study was conducted during the turnover period after the Chinese New Year (i.e., when turnover is the highest). We targeted the population with the most frequent turnover rate (i.e., individuals between 20 and 39 years old), and we divided the population into high and low job fluctuation. The survey was based on hierarchical random sampling. Assessment of reliability and validity indices.

The reliability measurement of this study was based on Cronbach's α, and the reliability coefficient needed to be greater than .7. To evaluate validity, we extracted the main information using factor analysis to replace the original variables and avoid colinearity. We then chose variables according to reliability, factor loading > .5 and explanatory variance, and we compared the variables with previous studies to validate the construct validity of the present study.

3.1.2.1. Questionnaire design. The present study used six measuring instruments: (1) the Minnesota Satisfaction Questionnaires, which aimed to measure respondent job satisfaction; (2) an organizational commitment scale, which was based on a local scale developed by the Department and Graduate Institute of Business Administration and measured local domestic organizational commitment; (3) a supervisor commitment scale, which was based on a local scale developed by the Department and Graduate Institute of Business Administration and measured domestic supervisor loyalty; (4) a job stress scale; (5) a turnover tendency scale; (6) and an organizational politics perception scale. This study extracted 24 variables using factor analysis, and the data were normalized according to four demographic variables. The selected variables are shown in Fig. 3.

3.2. Sample cluster analysis

This study categorized samples into high and low job fluctuation and then clustered the data using a SOM with artificial neural networks (SOM + BPN). This study evaluated the appropriate number of clustering by three indices.

(1) A practical construct that explains varimax with the least number of clusters.
(2) The frequency of learning cycles.
(3) Root Mean Square Error (RMSE).

This study aimed to find the individual characteristics of the highest turnover tendency clusters from various variables. To investigate the clustering effect, this research adopted the SOM + BPN, which generally included hierarchical and non-hierarchical clustering. This research further validated the differences among variables generated from ANN classification. The following is a brief introduction of the main steps of analysis.

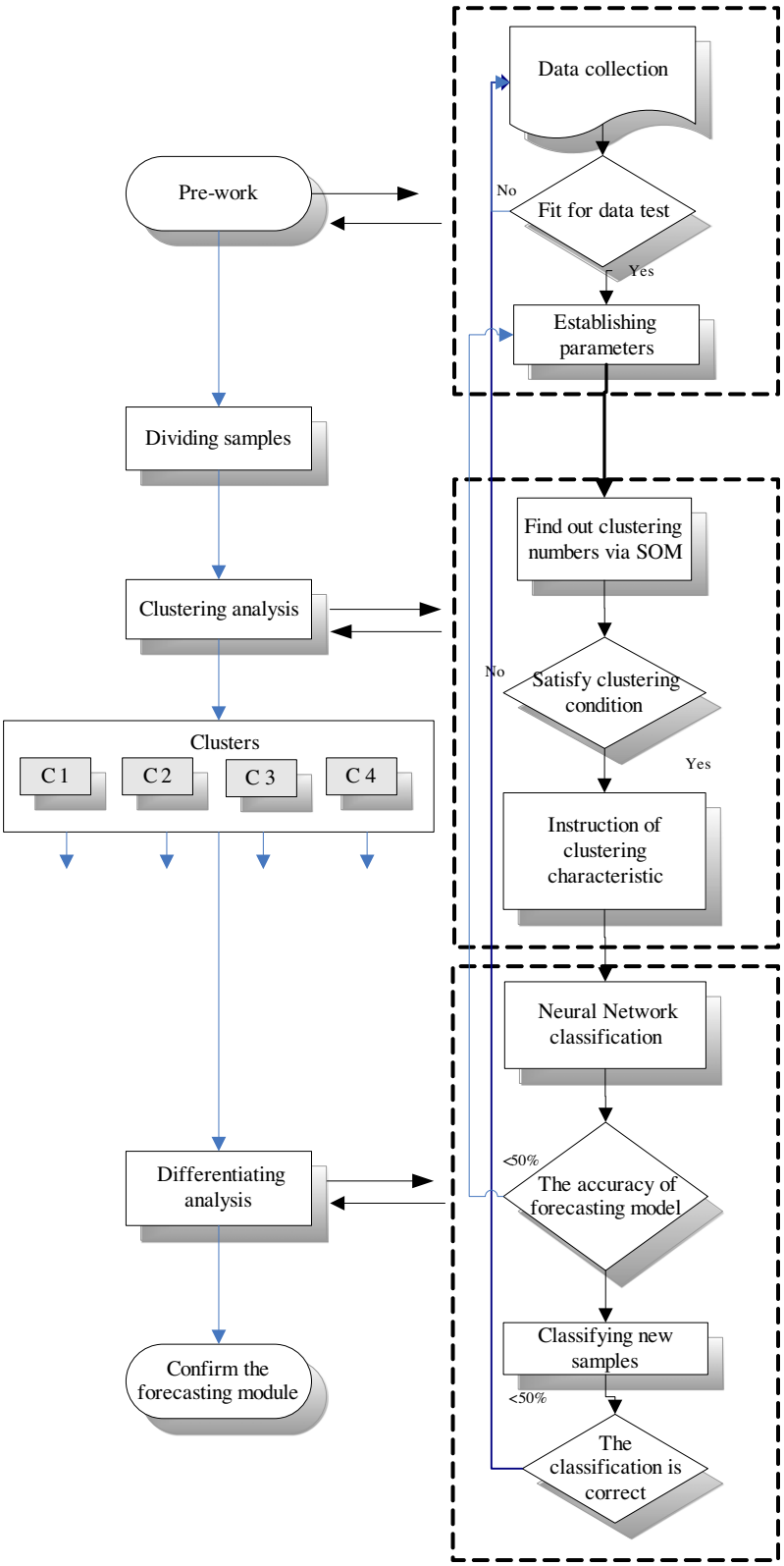

Fig. 1. Study flow chart.

3.2.1. The SOM clustering process

The present study utilized software Statistical 7.0 to develop our major SOM executive tools. The following steps describe the detailed process:

Step 1: Initialize each neuron to show the turnover intention weight, $w_i = [w_{i1}, w_{i2}, w_{i3}, \ldots, w_{ij}]^T \in j$. In this research, neuron weights were initialized by drawing random samples from the input dataset.

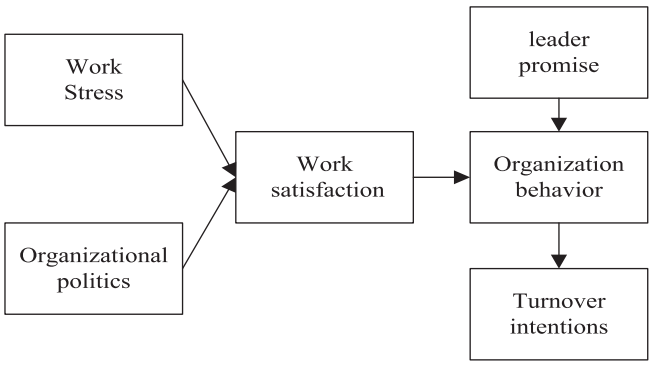

Fig. 2. Defining the research model according to turnover trends.

Step 2: Present an input pattern $x = [x_1, x_2, x_3, \ldots, x_j]^T \in j$. In this case, the input pattern was a series of variables representing new turnover intention distances for pattern x and each neuron weight (w_i). We also wanted to identify the best-matching unit (w_C),

$$\|x - w_c\| = \min\{\|x - w_i\|\} \quad (1)$$

Step 3: Adjust the weights of the winning neuron (c) and all neighboring units.

$$w_i(t+1) = w_i(t) + h_{ci}(t)[x(t) - w_i(t)], \quad (2)$$

where w_j is the index of the neighboring neuron, and t is an integer to indicate discrete time. The neighborhood kernel $h_{ci}(t)$ is a function of time, and the distance between neighbor neuron i and winning neuron $ch_{ci}(t)$ defines the region of influence that the input pattern has on the SOM. The neighborhood kernel in SOM consists of two parts (Kuo, Ho, & Hu, 2001): the neighborhood function, $h(\|\cdot\|, t)$, and the learning rate function, $a(t)$, which are shown in Eq. (3).

$$h_{ci}(t) = h(\|r_c - r\|_i, t)a(t). \quad (3)$$

In Eq. (3), r is the location of the neuron on two-dimensional map grids. In this work, we used the Gaussian neighborhood function. The learning rate function, $a(t)$, is a decreasing function of time. The final form of the neighborhood kernel under the Gaussian function is defined by the following equation:

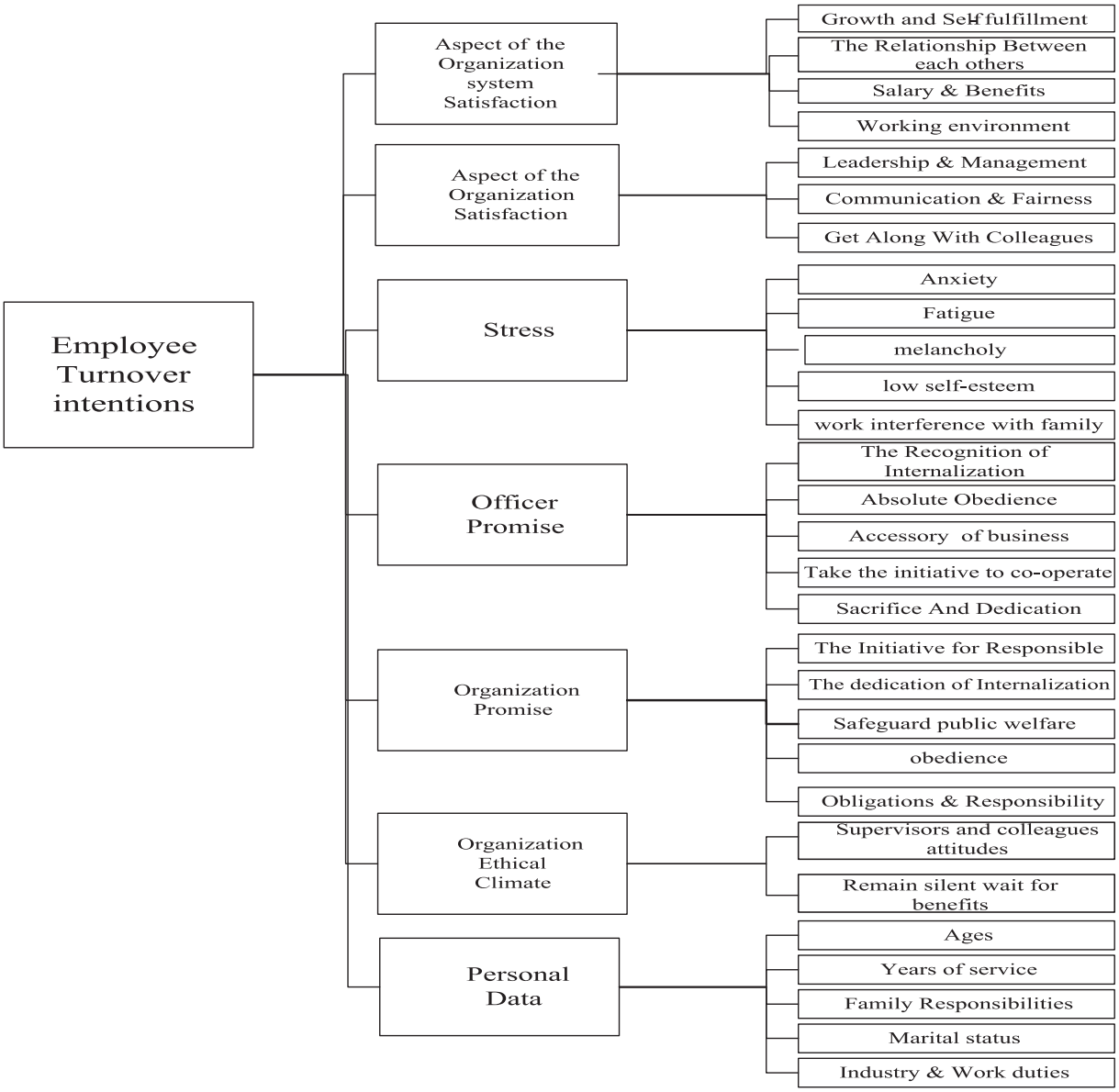

Fig. 3. Questionnaire design flow chart.

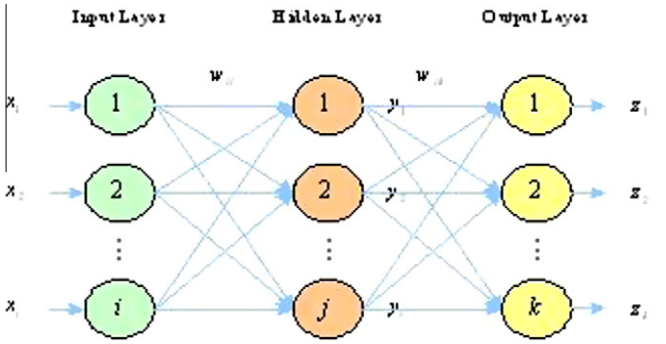

Fig. 4. The structure of a back-propagation neural network.

$$h_{ci}(t) = \exp\left(\frac{\|r_c - r_i\|}{2\sigma^2(t)}\right) a(t) \quad (4)$$

where $a(t)$ defines the width of the kernel.

Step 4: Repeat steps 2 and 3 until the convergence criterion is satisfied. Through a SOM learning process, this research uses SOM method to judge the best-clustering groups.

3.2.2. The neural network clustering method

Artificial neural networks is one kind of machine learning system that has been shown to be effective at approximating complex nonlinear functions Berardi and Zhang (1999). For classification tasks, complex nonlinear functions represent the shape of the partitions between classes.

Artificial neural networks traditionally consist of nodes in many layers. Data are input through the network layer, which is called the input layer, processed through one or more hidden layers, and output through the output layer. Nodes in each layer are connected to nodes in the next layer. Each hidden layer and output layer node transform data from upper-level nodes and applies an activation function. The activation function produces a value of either 1 or 0 (i.e., 1 turn the node on and 0 turn the node off) and passes the value to connected nodes in the subsequent layer. Connections between nodes are weighted so that nodes may include a bias. For example, if the three nodes x_1, x_2 and x_3 feed into a fourth node, x_4, with a bias (θ) via connections weighted with w_1, w_2, and w_3, the value that x passes to the next layer would be determined by $x = f(\sum ix_i w_i + \theta)$ (Hart, 1992). A detailed depiction of this process is shown in Fig. 4.

The values of the weights and biases are determined through a training process. Data are fed forward through the network, and connection weights and node biases are adjusted through a back-propagation (BP) algorithm. This process of feed-forward, back-propagation is repeated until the output reaches a desired accuracy or a given number of training cycles has been completed.

For classification, the size of the input layer is usually the number of attributes of the dataset, and the size of the output layer is the number of classes. When there are only two classes, a single output layer may be used with specific target values, which indicate the class of the sample. The size and number of hidden layers are flexible, which is a drawback of BPN. Although guidelines exist to help determine the numbers of nodes and hidden layers that should be used, the architecture is largely determined through experimentation.

The present study combined these two methods (SOM + BPN) to derive the best forecasting results.

Table 3
Parameter settings from the SOM.

Parameter	Setting
Input levels	28
Output levels	1
Hidden levels	1
Learning rule	Delta rule
Transform rule	Sigmoid

Table 4
SOM clustering convergence speed analysis.

Clustering groups	Learning times	RMSE value	Convergence speed
$C = 2$	79,000	0.00900	
$C = 3$	79,000	0.00800	
$C = 4$	52,000	0.00850	Best
$C = 5$	80,000	0.00730	
$C = 6$	79,000	0.00720	

4. Analysis results

This research distributed 605 questionnaires to Taiwanese employees of different ages between January 1, 2009 and April 14, 2009. There were 511 returned questionnaires, which was a return rate of 85%. After eliminating the questionnaires with missing answers or significantly illogical responses, 421 valid questionnaires remained. Therefore, the valid return rate was 70%. This chapter shows the validation of the research samples, the reliability test and the variables. In addition, this chapter describes the clustering results from SOM + BPN, the predictor model and the validated clustering performance.

Table 2
Input variables.

Input variable number	Input variable name	Input variable number	Input variable name
X1	Growth and self-fulfillment	X15	Accessory of business
X2	The relationship between one another	X16	Take the initiative to co-operate
X3	Salary and benefits	X17	Sacrifice and dedication
X4	Working environment	X18	The initiative to take responsibility
X5	Leadership and management	X19	The dedication of internalization
X6	Communication and fairness	X20	Safeguard public welfare
X7	Getting along with colleagues	X21	Obedience
X8	Anxiety	X22	Obligations and responsibilities
X9	Fatigue	X23	Supervisors and colleague attitudes
X10	Melancholy	X24	Remaining silent while waiting for benefits
X11	Low self-esteem	X25	Age
X12	Work interference with family	X26	Years of service
X13	The recognition of internalization	X27	Family responsibilities
X14	Absolute obedience	X28	Marital status

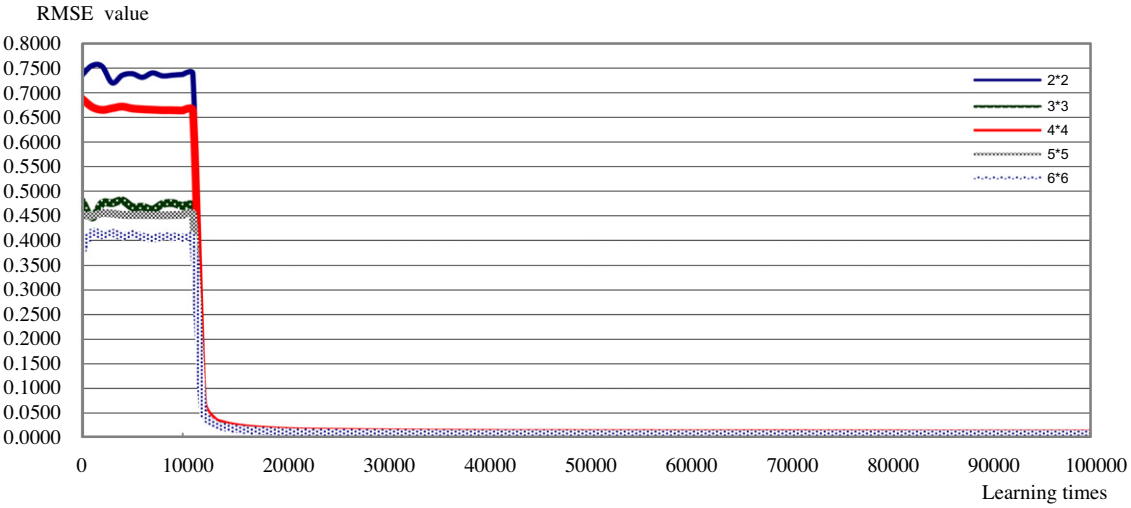

Fig. 5. The learning curves for the SOM for 2–6 clusters.

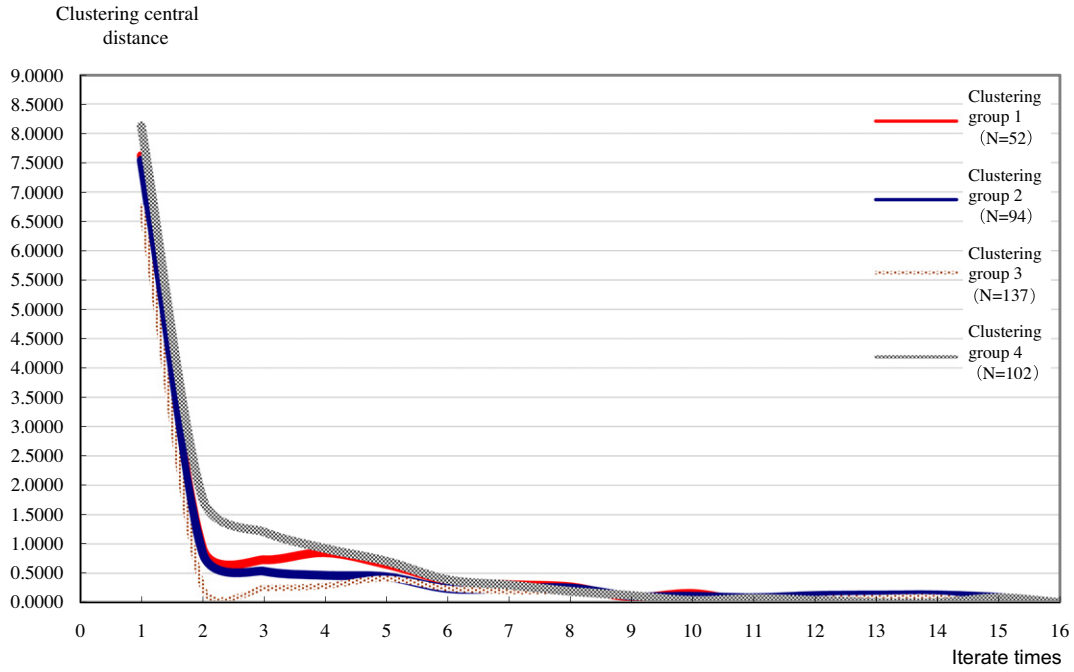

Fig. 6. The learning curves of the SOM + BPN clustering analysis for four groups.

4.1. The SOM clustering results

The SOM clustering method has been used for the initial classification of clustering groups. In the present study, we randomly selected 36 samples from the 421 samples; these 36 samples have been consider test examples. Other 385 samples were used in the SOM model for cluster analysis. After identifying the best-clustering groups (shown in Tables 3 and 4, Fig. 5), the K-means clustering method was applied to cluster all data sets. Following the suggestions outlined in Section 3, this research focused on 28 variables, which are shown in Table 2. Focusing on clustering efficiency, this research chose to use the RMSE value for the SOM clustering major index, and a clustering range from 2 to 6 was determined to the best one.

4.2. SOM + BPN clustering result

After we obtained the initial clustering group from the SOM, we used ANN and BPN classification for all data clustering. As shown in Fig. 6, convergence was completed in 16 iterations, which returned the best results.

Through two-phase SOM + BPN clustering, this approach successfully clustered all data into four groups: a high turnover tendency cluster, a medium turnover tendency cluster, a second-lowest turnover tendency cluster and a lowest turnover tendency cluster. The distribution of the four clusters is shown in Fig. 7. After completely distinguishing the four groups, this research randomly chose four samples to test the accuracy rate of the hybrid model.

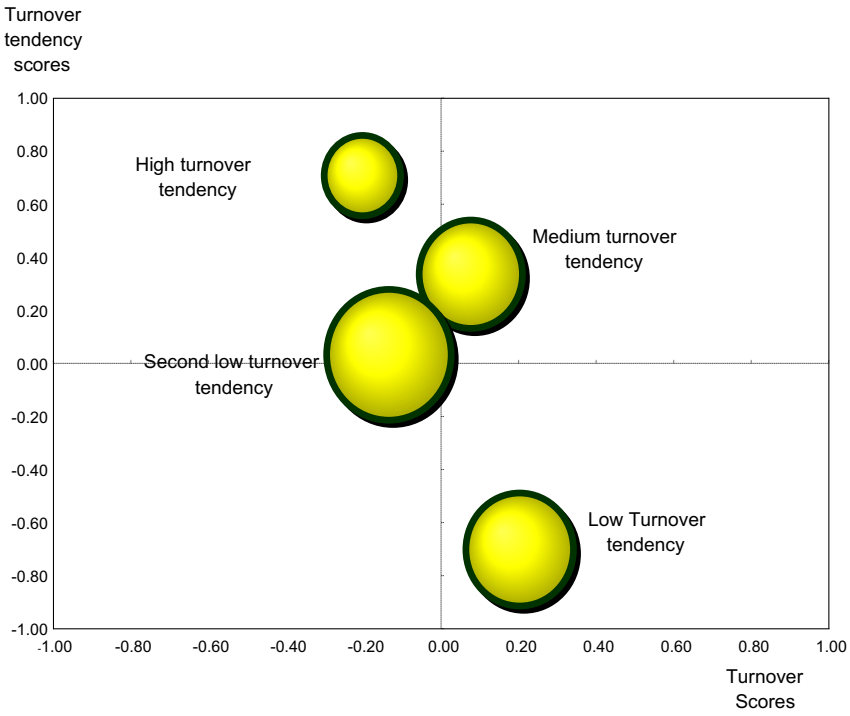

Fig. 7. The two-phase clustering result (i.e., the turnover tendency).

Table 5
Thirty-six test data forecasting results.

Test data	Clustering groups				Evaluation	
	1	2	3	4		
1	−4.003	−4.453	−2.632	−4.195	3	3
2	−6.358	−0.533	−4.379	−4.495	1	2
3	−10.131	−1.276	−3.820	−2.637	2	2
4	−10.574	−1.672	1.559	−8.895	3	3
5	−7.289	−7.429	−1.756	−1.005	4	4
6	11.278	−3.237	−5.120	−11.033	1	1
7	7.598	−6.674	−3.898	−6.959	1	1
8	−9.820	−0.662	−2.994	−4.473	2	2
9	−11.547	1.861	−0.748	−8.446	1	2
10	−11.430	−6.079	−2.277	0.444	4	4
11	−11.553	−4.284	−2.651	−0.557	4	4
12	−7.173	−7.807	1.066	−4.206	3	3
13	−12.245	−3.574	−1.152	−2.676	3	3
14	−8.135	−4.933	−1.165	−3.445	3	3
15	−1.225	−5.550	−3.756	−3.290	1	1
16	3.897	−5.824	−4.361	−4.855	1	1
17	−4.069	−1.649	−5.494	−3.266	2	2
18	−0.988	0.550	−6.710	−5.541	2	2
19	−9.142	1.260	−5.720	−2.881	2	2
20	−10.428	−2.277	−3.727	−1.679	4	4
21	−7.219	−4.138	−4.358	−0.413	4	4
22	−7.848	−3.252	−0.891	−5.433	3	3
23	−5.830	−2.348	0.750	−9.130	3	3
24	0.302	−4.912	−2.402	−6.115	1	1
25	−0.474	−5.719	−5.349	−2.049	1	1
26	−4.559	−0.268	−2.972	−7.558	2	2
27	−5.418	−0.165	−5.215	−4.348	2	2
28	−11.298	0.020	−9.489	4.074	4	4
29	−12.260	−1.690	−7.733	3.442	3	4
30	−7.532	0.989	−7.978	−0.657	2	2
31	−5.957	−11.819	1.595	−1.639	3	3
32	−9.301	−4.400	−4.758	0.721	4	4
33	−11.886	−1.960	−6.489	2.192	4	4
34	−7.184	−0.889	−5.679	−1.977	2	2
35	−5.249	−1.170	−8.365	0.426	4	4
36	−5.050	−2.734	−6.725	−0.473	4	4

Table 6
Overall comparisons for BPN, K-means, and SOM + BPN.

Method	K-means	BPN	SOM + BPN
Accuracy of forecasting results (%)	63.5	87.2	92

Table 5 shows the results of the test data forecasting using the hybrid model, the gray rows means wrong classification in this methodology. Which exhibited an accuracy rate of greater than 92%. In addition, the hybrid model was superior to the traditional models, including the K-means model and the single ANN model. Detailed results are shown in Table 6.

5. Conclusions

Previous studies on turnover tended to explore the factors behind turnover, but they rarely predicted turnover inclinations. Therefore, they could not be used to immediately and effectively avoid the loss of key employees. This study solved this problem by using cluster analysis in data mining. According to the characteristics of the clusters, we found that employees value certain factors, such as the internalized identifications of supervisors, business support, leadership, management, growth and self-realization. In terms of turnover tendencies, the influence of supervisor commitment (or loyalty) was more significant than organizational commitment, which was distinct from other turnover research results on Western countries. Managers should not assume that the employees are less likely to leave when they actively fulfill their duties or when companies increase salary and benefits. In addition, when companies strictly follow procedure and rules and neglect the human side of the workplace, supervisors can accumulate loyal subordinates and then encourage all of them to leave for another firm. Clearly, it is much more important to cultivate loyal employees than abide by each and every rule.

Due to time and cost considerations, this study still can improve forecasting efficiency and research application by use

other methods. Thus, we propose that future scholars develop studies that do not consider turnover rates based on dichotomy. Thus, we might include fuzzy theory and multiple criteria decision-making to find additional influential factors with respect to turnover. To more effectively filter out the valid questionnaires, we might include traps and evaluate valid questionnaires through the use of a design program, which would save a significant amount of time.

References

Berardi, V. L., & Zhang, G. P. (1999). The effect of misclassification costs on neural network classifiers. *Decision Sciences, 30*(3), 659–682.

Dalton, D. R., Todor, W. D., & Krackhardt, D. M. (1982). Turnover overstated: The functional taxonomy. *Academy of Management Review, 7*, 118–119.

Hao, C. C., Jung, H. C., & Yenhui, O. (2009). A study of the critical factors of the job involvement of financial service personnel after financial tsunami: Take developing market (Taiwan) for example. *African Journal of Business Management, 3*(12), 798–806.

Hart, S. (1992). Games in extensive and strategic forms. In R. J. Aumann & S. Hart (Eds.), *Handbook of game theory with economic applications*. Amsterdam: North-Holland.

Kraut, A. I. (1975). Predicting turnover of employees form measured job attitudes. *Organizational Behavior and Human Performance, 13*, 233–243.

Kuo, R. J., Ho, L. M., & Hu, C. M. (2001). Integration of self-organizing feature map and K-means algorithm for marketing segmentation. *Journal of Computers and Operation Research*.

Michaels, C. E., & Spector, P. E. (1982). Causes of employee turnover: A test of The Mobley, Griffrth, Hand, and Meglino model. *Journal of Applied Psychology, 67*(1), 53–59.

Miller, H. E., Miller, H. E., Katerberg & Hulin, C. L. (1979). Evaluation of the Mobley, Horner and Hollingsworth model of employee turnover. *Journal of Applied Psychology, 64*, 509–517.

Mobley, W. H. (1977). Intermediate linkages in the relationship between job satisfaction and employee turnover. *Journal of Applied Psychology, 62*(2), 237–240.

Mobley, W. H., Griffeth, R. W., Hand, H. H., & Meglino, B. M. (1979). Review and conceptual analysis of the employee turnover process. *Psychological Bulletin, 86*(3), 493–522.

Mobley, W. H., Horner, S. O., & Hollingsworth, A. T. (1978). An evaluation of precursors of hospital employee turnover. *Journal of Applied Psychology, 63*(4), 408–414.

Newman, J. E. (1974). Predicting absenteeism and turnover: A field comparison of Fishbein's mold and traditional job attitude measures. *Journal of Applied Psychology, 59*, 610–615.

Porter, L. W., & Steers, R. M. (1973). Organizational, work, and personal factors in employee turnover and absenteeism. *Psychological Bulletin, 80*(2), 151–176.

Samuel, M. O., & Chipunza, C. (2009). Employee retention and turnover: Using motivational variables as a panacea. *African Journal of Business Management, 3*(9), 410–415.