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.