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.
E-mail: 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 subset selection used in this study combined Taguchi method and Nearest Neighbor Classification Rules to select 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 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 subset selection (80.93%). The new 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; 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 (Porter, 1974) 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, machine learning, and statistical pattern recognition [1]. 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] 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 [32] and data mining [32] 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].

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

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.

3-1-1 Purposes of Feature Selection

Feature subset selection offers many advantages for pattern classification. Firstly, the cost of gathering training or unseen instances can be reduced. Secondly, pattern classification or machine learning models can be constructed faster. Furthermore, the performance (i.e., classification accuracy) and the comprehensibility of the learning models can be improved.

3-2-2 Feature selection consists of four important steps

Feature subset selection is generally carried out in four steps [24-26]. (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 or ranking method to rank or 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].

4 A Novel Approach to Hybrid Feature Selection Model

4-1 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.

This 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:

The author assumed that there are m training samples, representing as V= {v1, v2, ..., vm}.
Each sample had $n$ categories, represented as $F = (f_1, f_2, \ldots, f_n)$. 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 $S_j$.

**Step 2.** The nearest neighborhood rule and Leave-one-out cross-validation were used for each feature subset, $S_j$ to find their average accuracy, represented as $ACC(V, S_j)$. $ACC (V, S_j)$ 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 subset. 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 for feature selection estimation.

4-2 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 $S_j$ for each experiment $j$, labeled as $ACC (V, S_j)$. 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 $S_j$, 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 are 79.93% 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.

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.

Table 5-1  features related to turnover categories

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

6 Conclusions

The hybrid model suggested in this study was the combination with Taguchi Methods and the nearest neighborhood rule. 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 neighborhood 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.

References:


