Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 280–290
Expert Systems with Applications
www.elsevier.com/locate/eswa

Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry

Chen-Fu Chien [a,*], Li-Fei Chen [a,b]

[a] *Department of Industrial Engineering and Engineering Management, National Tsing Hua University, 101 Section 2 Kuang Fu Road, Hsinchu 300, Taiwan, ROC*

[b] *Department of Industrial Engineering and Management, Tahua Institute of Technology, 1 Ta-Hwa Road, Chung-Lin, Hsinchu 307, Taiwan, ROC*

Abstract

The quality of human capital is crucial for high-tech companies to maintain competitive advantages in knowledge economy era. However, high-technology companies suffering from high turnover rates often find it hard to recruit the right talents. In addition to conventional human resource management approaches, there is an urgent need to develop effective personnel selection mechanism to find the talents who are the most suitable to their own organizations. This study aims to fill the gap by developing a data mining framework based on decision tree and association rules to generate useful rules for personnel selection. The results can provide decision rules relating personnel information with work performance and retention. An empirical study was conducted in a semiconductor company to support their hiring decision for indirect labors including engineers and managers with different job functions. The results demonstrated the practical viability of this approach. Moreover, based on discussions among domain experts and data miner, specific recruitment and human resource management strategies were created from the results.

Keywords: Personnel selection; Human capital; Data mining; Decision tree; Semiconductor industry

1. Introduction

Human capital is one of the core competences for high-tech companies to maintain their competitive advantages in the knowledge economy. Personnel recruitment and selection directly affect the quality of employees. Hence, various studies have been conducted on resumes, interviews, assessment centers, job knowledge tests, work sample tests, cognitive tests, and personality tests in human resource management to help organizations make better personnel selection decisions. Indeed, the existing selection approaches focus on work and job analysis that are defined via specific tasks and duties based on their static properties.

However, owing to the changing nature of knowledge workers in high-tech industry, jobs cannot be easily delineated especially for jobs in the management level. As globalization and technology advance, cross-functional tasks are also increased while new jobs are also constantly created. The requirements of personnel quality in high-technology companies are increasingly strict, while the work processes in these companies are becoming diversified and complicated. Thus, the conventional personnel selection approaches that are developed on the basis of static job characteristics will no longer suffice (Lievens, Van Dam, & Anderson, 2002). In order to find the right people to do the right things for the right jobs, developing effective selection approaches is very critical.

A high-tech industry such as semiconductor industry has many unique or unusual characteristics including complex and highly uncertain manufacturing processes, short product life cycles, low yield problems, and difficulties in

[*] Corresponding author. Tel.: +886 3 5742648; fax: +886 3 5722685. *E-mail address:* cfchien@mx.nthu.edu.tw (C.-F. Chien).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.003

acquiring human capital (Chien & Wu, 2003; Sattler & Sohoni, 1999). Thus, the quality of their human resource is very crucial in increasing their competitiveness. In addition, Appleyard and Brown (2001) analyzed the firm-level data from semiconductor manufacturers in the United States, Asia, and Europe and found that engineers play important and growing roles in creating high-performance semiconductor factories. Nevertheless, semiconductor companies, as well as other high-technology companies, often suffer from high turnover rates and difficulties in recruiting the right talents. In order to attract good applicants, companies provide attractive compensation and welfare benefits. However, despite the willingness of many companies to do all that they can to recruit the best people, they usually have difficulties at the selection stage in predicting which applicants would have better work performance and would have longer service time after they are hired. Therefore, selecting the right engineers who can demonstrate the best performance and who will stay with the company for a long time is of great urgency for every high-technology company.

Recently, owing to the advancements in information technology, researchers have developed decision support systems and expert systems to improve the outcomes of human resource management. In particular, data mining is recognized as one of the most salient topics. Data mining refers to the extraction of useful patterns or rules from a large database through an automatic or semi-automatic exploration and analysis of data (Berry & Linoff, 1997; Chen, Han, & Yu, 1996). With the help of data mining techniques, computers are no longer limited to passively storing or collecting data. They can also help the users to actively excerpt the key points from huge amounts of data, and make use of analysis or prediction. Data mining techniques have been widely applied in many fields and have exhibited outstanding results. However, the applications of data mining in the semiconductor industry are mostly related to engineering data analysis and yield enhancement (Braha & Shmilovici, 2002; Kusiak, 2001; Chien, Hsiao, & Wang, 2004; Chien, Wang, & Cheng, 2007). Little research has been done in human resource management.

This study aims to develop a data mining framework for personnel selection to explore the association rules between personnel characteristics and work behaviors, including work performance and retention. An empirical study for indirect labor (IDL) including engineers with different job functions in one of the world largest semiconductor foundry company located in the Hsinchu Science Park in Taiwan is studied to demonstrate the validity of this approach. In particular, we employ decision tree analysis to discover latent knowledge and extract the rules to assist in personnel selection decisions. Furthermore, using the information gathered, domain experts from this company can also generate recruiting and human resource management strategies. Some of the findings have been implemented in this company and the results have shown the practical viability of this approach.

2. Fundamentals

2.1. Personnel selection

Personnel selection plays a decisive role in human resource management in which it will determine the input quality of personnel. Researchers (Borman, Hanson, & Hedge, 1997; Robertson & Smith, 2001) reviewed the personnel selection studies and found that the important issues including change in organizations, change in work, change in personnel, change in the society, change of laws, and change in marketing have influenced personnel selection and recruiting. Hough and Oswald (2000) also reviewed personnel selection studies from 1995 through 1999 and concluded that the nature and analysis of work behavior are changing and hence affecting personnel selection practices. Lievens et al. (2002) identified challenges in personnel selection including labor market shortages, technological developments, applicant perception of selection procedures, and construct-driven approaches.

Meanwhile, advancements in information technology are also affecting personal selection as well as human resource management (Beckers & Bsat, 2002; Kovach & Cathcart, 1999; Liao, 2003). The applications of expert systems or decision support systems on personnel selection and recruitment are increasing (e.g., Hooper, Galvin, Kilmer, & Liebowitz, 1998; Nussbaum et al., 1999). However, little research has employed data mining approaches for personnel selection as the present study does.

2.2. Data mining

Date mining methodologies have been developed for exploration and analysis, by automatic or semi-automatic means, of large quantities of data to discover meaningful patterns and rules. Indeed, such data including personnel data can provide a rich resource for knowledge discovery and decision support. Therefore, data mining is discovery-driven not assumption-driven. Data mining involves various techniques including statistics, neural networks, decision tree, genetic algorithm, and visualization techniques that have been developed over the years.

Data mining problems are generally categorized as association, clustering, classification, and prediction (Fayyad, Piatesky-Shapiro, & Smyth, 1996; Fu, 1997; Han & Kamber, 2001). Association is the discovery of association rules showing attribute-value conditions that occur frequently together in a given dataset. Clustering is the process of dividing a dataset into several clusters in which the intraclass similarity is maximized while the inter-class similarity is minimized. Classification derives a function or model that identifies the categorical class of an object based on its attributes. Prediction is a model that predicts a continuous value or future data trends.

Data mining has been applied in many fields such as marketing, finance, banking, manufacturing, health care, customer relationship management, failure detection and prediction, and organization learning (e.g., Chien, Chen, & Lin, 2002; Peng & Chien, 2003; Peng, Chien, & Tseng, 2004; Shiue & Su, 2003; Shaw, Subramaniam, Tan, & Welge, 2001; Wei & Chiu, 2002; Wu, Kao, Su, & Wu, 2005). However, its application in human resource management is rare. In particular, Cho and Ngai (2003) used data mining to develop a decision support system to predict the length of service, sales premiums, and persistence indices of insurance agents. The authors (Chien, Wang, & Chen, 2005) also employed data mining to analyze misoperation behaviors of operators.

2.3. Decision tree

Decision tree is a data mining approach that is often used for classification and prediction. Although other methodologies such as neutral network can also be used for classification, decision tree has the advantages of easy interpretation and understanding for the decision makers to compare with their domain knowledge for validation and justify their decisions. In addition, decision trees can analyze various data without requiring the assumptions about the underlying distribution. Thus, the proposed approach is based on decision tree for human resource data mining to generate rules for personnel selection.

Decision trees are usually presented in a tree structure, with leaves and stems. The hierarchical structure of decision trees could analyze different levels of factors. Every leaf reveals the classification result, while the stems indicate the conditions of the attributes. Given a set of training instances, including input variables and a corresponding output variable, a decision tree can be constructed depending on certain learning strategies to sort the variables into classes and provide the inductive rules. In particular, several algorithms such as CART (Breiman, Friedman, Olshen, & Stone, 1984), CHAID (Kass, 1980), ID3 (Quinlan, 1986), and C4.5 (Quinlan, 1993) have been developed for decision tree induction. In particular, CHAID (i.e., Chi-squared automatic interaction detection) is a non-binary decision tree that is designed specifically to deal with categorical variables and can determine the best multi-way partitions of the data on the basis of significance tests (Kass, 1980). CART (i.e., classification and regression tree) is a binary decision tree with the Gini index of diversity as the splitting criterion, and pruning by minimizing the true misclassification error estimate (Breiman et al., 1984). CART can deal with categorical and continuous variable. C4.5 is a variant and extension of a well-known decision tree algorithm, ID3 (Quinlan, 1993). The splitting criterion of C4.5 algorithm is gain ratio that expresses the proportion on information generated by a split. The error-based pruning is used for pruning in C4.5. In general, the objective of these algorithms is to maximize the distance between classes. These algorithms can be distinguished in terms of different distance measurements, pruning methods, missing value disposition, the number of branches at each node, and the data type they could handle. Table 1 summarizes the comparison of various induction algorithms.

3. Approach

This research constructed a framework for human resource data mining to explore the relationships between personnel profiles and work behaviors. Through the proposed methodology, hidden information could be extracted from large volumes of personnel data and thus the decision makers can have a better understanding and visualization of such latent knowledge. The discovered rules can be used to identify effective sourcing channels to find right talents and develop selection criteria. Fig. 1 shows the framework with the following steps:

(1) Problem definition and objective structuring:
The first step in data mining is to understand and define the right problem and specify the objectives. Meanwhile, data miner should also equip themselves with domain knowledge to understand problem nature, which will greatly improve data mining effectiveness and efficiency. Indeed, human resource management activities are very complicated and thus few quantitative approaches have been employed in practice.

(2) Data collection and preparation:
Through understanding the sources and types of related data that can be gathered, collecting the right data is the basis of data mining. Indeed, human resource data usually stored in separate database

Table 1
A comparison of CART, CHAID, ID3, and C4.5

Algorithm	Authors	Data type	Tree-pruning methods	Number of branches at each node	Missing value methods	Split criteria
CHAID	Hartigan (1975)	Discrete	No pruning	Two or more	Missing value branch	P value for Chai-square test
CART	Breiman et al. (1984)	Discrete and continuous	Overall error rate	Two	Alternate/surrogate splits	Gini value entropy
ID3	Quinlan (1986)	Discrete	No pruning	Two or more	Cannot handle	Information gain
C4.5	Quinlan (1993)	Discrete and continuous	Estimated error rate	Two or more	Probability weight	Gain ratio

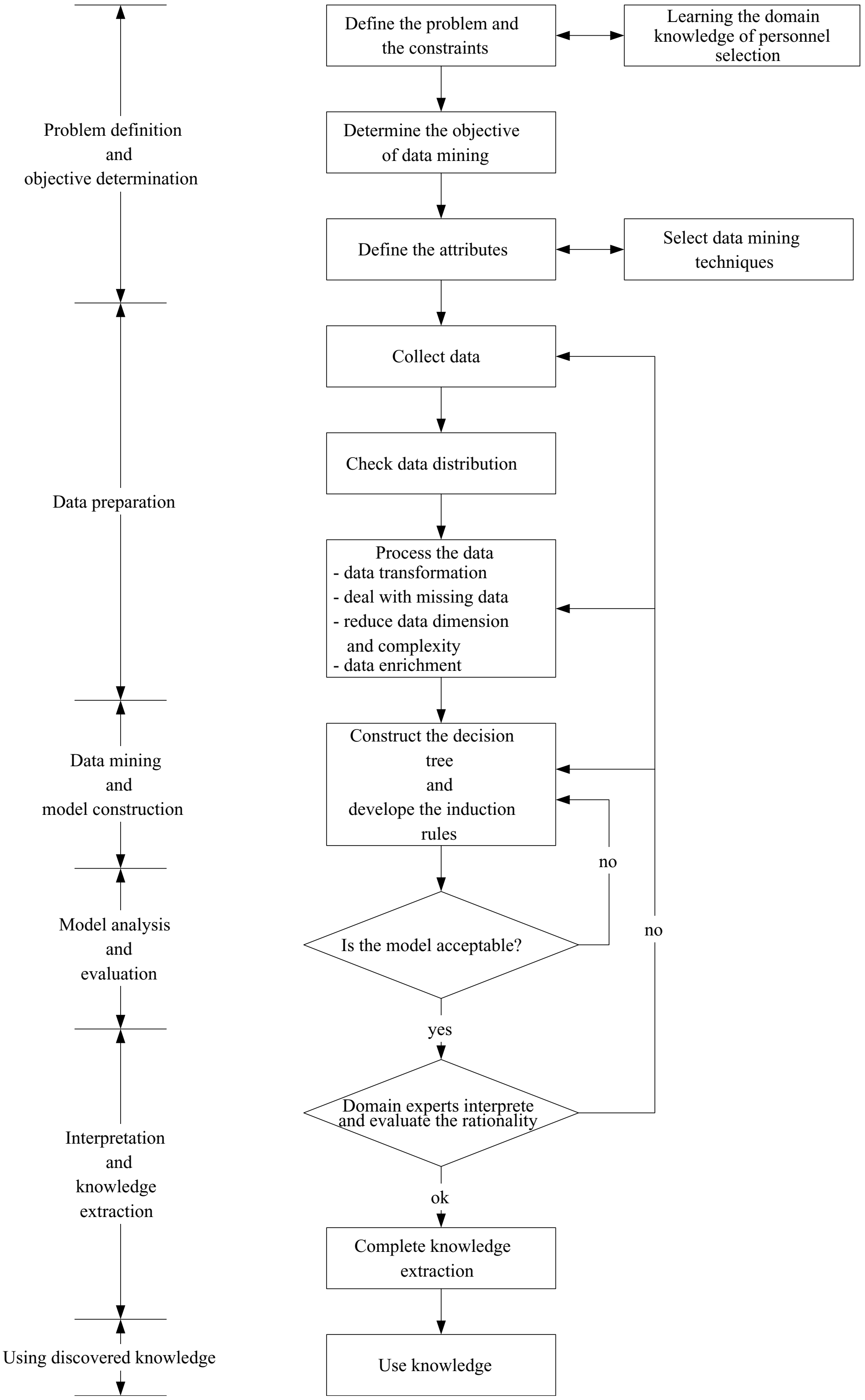
Problem definition
and
objective determination
Data preparation
Data mining
and
model construction
Model analysis
and
evaluation
Interpretation
and
knowledge
extraction
Using discovered knowledge
Define the problem and the constraints
Learning the domain knowledge of personnel selection
Determine the objective of data mining
Define the attributes
Select data mining techniques
Collect data
Check data distribution
Process the data
- data transformation
- deal with missing data
- reduce data dimension and complexity
- data enrichment
Construct the decision tree and develope the induction rules
no
no
Is the model acceptable?
yes
Domain experts interprete and evaluate the rationality
ok
Complete knowledge extraction
Use knowledge

Fig. 1. A data mining framework for personnel selection.

for privacy. The related data need to be combined and prepared before further analysis. However, the collected data often include noisy, missing and inconsistent data. Following Pyle (1999), data preparation processes that consist of checking the data distribution and outliers, dealing with empty or missing values, enriching data, and transforming data into analyzable formats were employed to improve data quality and to thus enable effective data mining.

(3) Data mining model construction:
The present problem on predicting work behaviors can be structured as a classification problem.

Considering the needs for result interpretation and rule justification, decision tree is employed for human resource data mining. Since most of the personnel data are categorical variables, CHAID is used to construct the tree via significant relationships among the variables for classification.

(4) Model analysis and evaluation:
The constructed model should be reviewed and evaluated before it can be used for decision support. To evaluate the model, we used lift as the criteria to assess the performance of the classification method and to select useful extracted rules. In addition, we also determine the acceptable sample size as a screening mechanism to keep only significant findings with sufficient data support. Moreover, in order to discover the most viable rules, we can choose different attributes to split the tree and hence form probable relationships among them.

(5) Interpretation and knowledge extraction:
Data mining results should be interpreted and assessed according to the experience and knowledge of domain experts in order to justify the meaning of extracted knowledge. For any unusual pattern or result which is counter common practice, a further study will be initiated to confirm its validity. Thus, the useful information or patterns can be extracted and summarized into decision support rules.

(6) Using discovered knowledge:
The discovered knowledge can be the basis for decision support to generate human management and personnel selection strategies. It can also be used to improve related management activities. Furthermore, since the empirical models derived from data mining have life cycle and thus need to be reviewed periodically to maintain its validity.

4. An empirical study

4.1. Background and significance

The case company was established in 1987 in the Hsinchu Science Park, Taiwan and is the global leader in semiconductor foundries. It has distinguished itself in the field by providing advanced wafer production processes and demonstrating incomparable manufacturing efficiency. Furthermore, this company has been continuously ranked as the most reputable enterprise, and it also offers the most attractive jobs to students. By the end of February 2005, its total workforce reached 18,570, including 1853 managers, 6715 professionals, 750 assistant engineers and clerical staff, and 9223 technicians. The company's employees are well educated including 2.4% with PhD degrees, 26.5% with Master degrees, 17.6% with university bachelor degrees, 23.9% with other college degrees, and 29.6% with high school diplomas. In addition, since this company was founded in 1987, most of the employees are quite young and their length of stay is not very long. The average age of her employees is 30.6 years old, while their average service year is 4.8 years. Among the professionals, engineers with different job grades including engineers, senior engineers, and chief engineers who were hired from 2001 to 2004 and thus have at least one record of annual performance evaluation were analyzed in this study.

Attracting and retaining the right and high-potential talents are the key objectives of a sound human resource strategy in knowledge economy. However, the recruiting department sustains heavy loads for hiring new employees due to low retention rate given high-pressure in high-tech industry. Furthermore, although every applicant has gone through long rounds of the selection processes before he or she could be hired, the line managers may still complain about high turnover rates and unacceptable performance. Thus, it is a challenge for many recruiters to predict the work behaviors of applicants by using limited information that could be gathered while in the selection stage. In addition, it is quite normal that applicants would show their best during the selection process, and this misleads the judgment of those involved in the recruitment.

There is a need for the case company to investigate the relationships between personnel data and their work behaviors so as to develop effective recruitment channels and the right screening criteria to identify the best talents in the selection stage for different job functions. Job performance, service length, and resignation are the major work behaviors considered and we focused on the subjects with outstanding or poor performance.

4.2. Problem definition and objective

In order to identify effective recruitment channels to access high-potential talents and design the appropriate screening criteria for selecting the right ones for different job functions, this study developed a data mining framework for analyzing human resource data, in which decision tree was employed to extract rules between applicants' profiles and their work behaviors. In other words, the objectives of the case company is to predict applicants' work behaviors including job performance and retention based on the inputs of profile attributes that can be obtained in the selection stage. These input variables include demographic data such as age, gender, marital status, educational background, work experience, and the recruitment channels such as internal or external. Fig. 2 illustrates the conceptual structure.

In particular, the data on the new employees hired from 2001 to 2004 for 19 job functions with specific job grades were collected and prepared for the present analysis. The total sample size is 5289. Since some job functions are sparsely populated, after discussions with domain experts, we focused on major function areas and thus the sample size was reduced to 3825. Furthermore, the engineers were categorized into five areas by combining functions with simi-

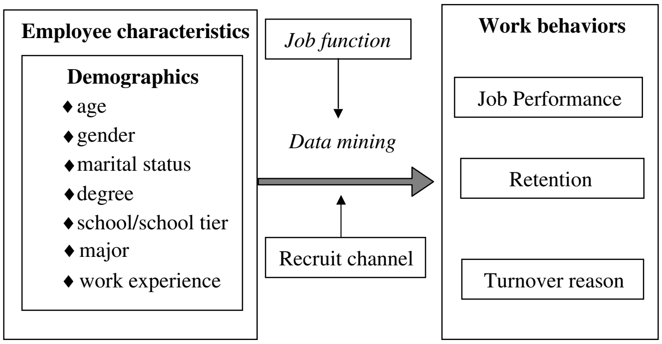

Fig. 2. Conceptual analysis framework.

larity in terms of desired characteristics. Tables 2 and 3 lists the general work descriptions of these five functions. For performance analysis, we focused on extracting information related to those who with job performances either “outstanding” (top 10%) or “improvement needed” (bottom 5%). As for retention analysis, 940 of the 3825 samples had already quit their jobs. According to the domain experience of human resource experts in the company, if one employee quits the job within three-month probation period, the recruitment process is considered failed and the investment of freshman training is wasted. On the other hand, if one employee quits the job within one year after he or she was hired, it is considered as a management problem. Thus, the characteristics of those who failed the probation and those who quit within one year are especially analyzed to uncover the relationship of personal profiles and job functions. We also considered the turnover reasons to help understand the factors affecting the retention rate to generate improvement strategy.

4.3. Data preparation

Because the collected personnel data are complex and often include noisy, missing, and inconsistent data, data preprocessing was conducted to improve the quality of the data. Then, the selected data are transformed into appropriate formats to support meaningful analysis. In particular, the target variables including job performance, retention, and turnover reasons are prepared as follows:

(1) Job performance: The company has established a performance management system to evaluate employees' performance. This annual performance review process provides a formal opportunity for each employee and the managers to discuss previous performance and set goals for the future development. They also rank employees' performance into three categories: outstanding (only top 10%), successful (85%), and improvement needed (bottom 5%). While the majority of employees were ranked as “successful”, the employees with the performances either “outstanding” or “improvement needed” should be differentiated.

(2) Retention (separate): Although it seems to be unavoidable that some new hired employee may not be able to fit in as expected, retention rate is a critical issue since this company invests a lot on her employee. Thus, this analysis was performed in two aspects, i.e., retention within three months due to failed recruiting process and retention within one year due to management and employee development failure.

(3) Turnover reasons: When an employee resigns, he or she would have an interview with the immediate supervisor and HR staff, in which 33 types of turnover reasons may be selected. The employee could choose up to three major reasons causing his or her resign, while the immediate supervisor can also choose one reason causing this employee to quit. Finally, after the interview, the HR staff would make

Table 3
The distribution of degree by different job functions

Function	Degree						Total	
	PhD & MS		BS		Others			
	N	%	N	%	N	%	N	%
A	491	31.1	1054	66.7	36	2.3	1581	100.0
B	887	87.6	125	12.4	0	0.0	1012	100.0
C	472	89.2	56	10.6	1	0.2	529	100.0
D	145	77.5	41	21.9	1	0.5	187	100.0
E	456	88.4	47	9.1	13	2.5	516	100.0
Total	2451	64.1	1323	34.6	49	1.3	3825	100.0

Table 2
The general work descriptions of different job functions

Function	General work description
A	The role is responsible for improving tool ability and performance, installing new tools, relocating used tools and checking tool performance, training and transferring company culture to newcomers, and troubleshooting tools
B	The role is responsible for process developing/tuning of lithography technology, and making the process successful after co-working with relative teams
C	The role is responsible for customer product handling, experiment lot planning and results analysis, and yield improvement
D	The role's responsibilities include device performance improvement/tuning, device yield improvement, new device definition/design, and cross-fab alignment
E	The role is responsible for the research and development of an advanced model design environment and methodology, support and solution of customers' design kit technical issues, and discussion with EDA tools vendors on any design kits and design flow problems

the judgment of the most possible turnover reason of the resigned employee. After discussing with domain experts, we categorized these data of turnover reasons into four categories including internal push factors, external pull factors, personnel issues, and company initiatives.

The following 8 profile variables can be collected during the selection stage and thus can be used as the predictors for this analysis. In particular

(1) Age: This attribute shows the age of the employee upon hiring. Rather than using the filled data, we used the variable derived from the birthday and hiring day for data integrity. Furthermore, we also transformed the data into meaningful categories.
(2) Gender: Female and male.
(3) Marital status: single, married without children, married with children, divorced with children.
(4) Experience: This attribute indicates previous work experiences. It is transformed into two categories: those who have more than one year of previous work experience that is denoted as ‘‘yes’’; otherwise, ‘‘no.’’
(5) Education: There were four categories of degrees: high school level and below, junior college degree, bachelor degree, and master and above.
(6) Major subjects: There were 52 different major subjects such as electrical engineering, material sciences, and industrial engineering in the original data set. However, only eleven majors have more than 50 samples and the other majors without sufficient samples are grouped into one category denoted as ‘‘others’’.
(7) School/School tiers: This variable denotes the school from which the employee was graduated from. Originally, there were 114 different universities in the data. However, this variable was transformed into four categories including three tiers of Taiwanese universities and fourth denoting the others.
(8) Recruitment channel: The recruitment channels include the internal channel and the external channel.

Although the input variables of age, gender, and marital status showed some interesting patterns, they were excluded them the analysis due to the concern of discrimination for hiring that may against the equity principle. Furthermore, we found most of the data were unbalanced distributed among the different instances of the variables. For example, most of the employees are majored in electrical engineering, male, graduated from tier one universities.

4.4. Data mining and model construction

We used CHAID as the data mining tool to explore the latent relationships among the input employee profiles and target variables of work behaviors such as job performance, retention, and turnover reasons. In particular, we investigated every possible tree structure by changing different splitting attributes from the root node to the leaves to find the potential relationships. For example, with job performance as the target, we conducted decision tree analysis with the input variables: degree, experience, recruitment source, and school tiers. We split the tree from the root node by using different input variables and hence to derive different decision tree structures. Fig. 3 illustrates one of the developed trees. Since the data set was unbalanced distributed, we explored alternative tree structures to seek the latent rules.

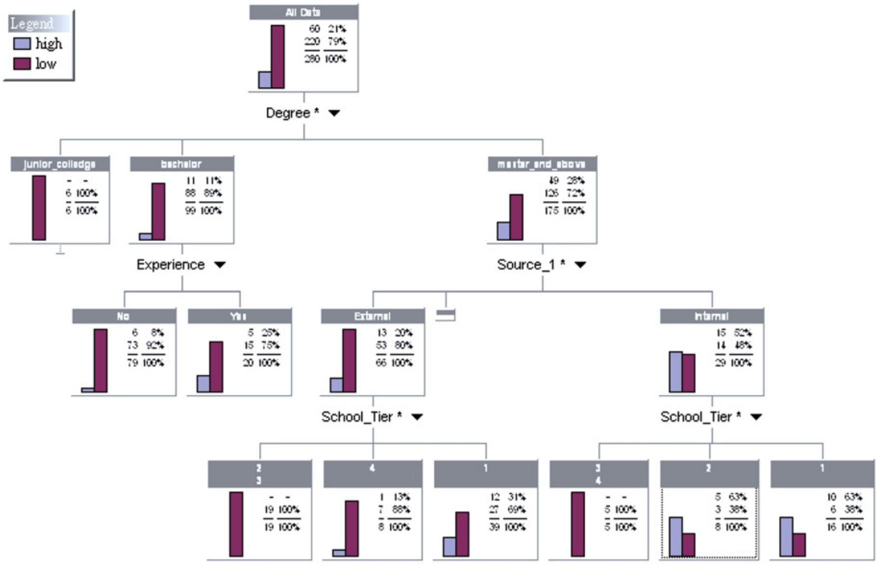

Fig. 3. Decision tree for predicting job performance.

4.5. Model analysis and evaluation

Furthermore, we employed the indexes, i.e., confidence and lift, to test the appropriateness of derived rules from constructed decision tree. In particular, confidence denotes the prediction accuracy that a subset can be categorized into a specific class. Lift is a ratio that is commonly used to assess the performance of the classification method. It indicates the change in concentration of a specific class when the rule is used to select a subset from the general population and thus the lift should be greater than one. That is,

$$\text{Confidence}_A(\text{Rule } i) = P(\text{class } A|\text{subset data selected by Rule } i) \qquad (1)$$

$$\text{Lift}_A(\text{Rule } i) = \frac{P(\text{target class } A|\text{subset } i)}{P(\text{target class } A|\text{population})} \qquad (2)$$

In addition, the tree was pruned so that the sample size of the leave node should not be less than 20 to have a significance level of supporting samples.

After the preliminary analysis, 50 rules associated with job performance and 16 rules associated with retention were discovered. Tables 4 and 5 list some of the derived rules. Table 6 shows the frequencies of each attribute that is connected to the predicted targets. As shown in the results, the variables of functions, school tiers, degree, and experience were the major attributes related to the predicted targets.

Table 4
Some examples of the job performance rules

No.	Rule	Lift
1	IF recruit channel = external THEN he/she will perform with a level of improvement needed. ($n = 95$; confidence = 84%)	1.06
2	IF experience = no THEN he/she will perform with a level of improvement needed. ($n = 214$; confidence = 83%)	1.05
3	IF degree = others THEN he/she will perform with a level of improvement needed. ($n = 105$; outstanding performance rate = 90%)	1.14
4	IF school tier = {2,4} THEN he/she will perform with a level of improvement needed. ($n = 149$; confidence = 86%)	1.09
5	IF degree = master's and above, recruitment channel = internal, and school tier = {1,2} THEN he/she will perform excellently. ($n = 24$; confidence = 63%)	3.00
6	IF degree = master's and above, recruitment channel = external, and school tier = {2,3,4} THEN he/she will perform with a level of improvement needed. ($n = 27$; confidence = 96%)	1.12

Table 5
Some examples of the derived rules for resignation

No.	Rule	Lift
1	IF function = C and experience = yes THEN he/she will quit within three months. ($n = 94$; resignation rate = 20%)	2.52
2	IF function = C and experience = no THEN he/she will not quit within three months. ($n = 245$; resignation rate = 94%)	1.02
3	IF function = B and recruit channel = internal THEN he/she will quit within one year. ($n = 75$; resignation rate = 31%)	1.15
4	IF function = B and recruit channel = external THEN he/she will not quit within three months. ($n = 208$; resignation rate = 79%)	1.08
5	IF function = C and experience = yes and recruit channel = external THEN he/she will quit within three months. ($n = 27$; resignation rate = 37%)	4.65

4.6. Interpretation and knowledge extraction

The results were then presented to a group of human resource experts for interpretation and discussions of potential usages of extracted rules. Finally, a total of 30 meaningful rules were chosen to develop the recruitment strategies. In particular, we discussed the proposed strategies according to recruitment channels, education, and work experience as follows:

(1) Recruiting channels:
The employees recruited from internal channels are more likely to exhibit better job performance than those who were recruited from external channels, though most of the existing employees were hired from external channels. Furthermore, although employee graduated from tier one university and employee graduated from tier two yet with master degree are generally considered with high-potential, we found internal hired were even significantly better than those from external channels. In addition, for the employees of function C, we found that the employees who had more than one year of previous work experience and were hired from external channels would be more likely to quit within three months

Table 6
Frequencies of each attribute that is related to the predicted targets

Target	Attribute						
	Functions	School/School tier	Recruitment channel	Degree	Experience	Job performance	Major
Job performance	14	26	26	22	20	0	2
Retention	22	8	8	4	4	6	0
Total	36	34	34	26	24	6	2

than those who were hired from internal channels. This finding was supported by domain knowledge since the employees recruited from internal channels usually stay longer, since the referrals should have shared more information about the job and the company. Thus, this company has adapted this finding and developed a strategy to raise a campaign for promoting the referrals by giving cash bonus to successful hiring.

Nevertheless, for the employees of function B, the employees hired from internal channels had not shown less proportion to quit within three months than those hired from external channels. The human resource staff compared their job contents with competitors and thus suggested the redesign of their role and responsibility of the employees in Function B.

(2) Education:

As shown in the results, the employees who graduated from first tier schools and those who with higher degrees were more likely to exhibit better performance in several dimensions. For example, the employees with a master or Ph.D. degree did show better performances than people with other degrees. Furthermore, for the employees with a higher degree and hired from external channels, those who were graduated from first tier schools would perform better than those who graduated from other schools. In addition, for the employees with a higher degree and hired from internal channels, those who graduated from the first or second tier schools would perform better than those who graduated from other schools.

However, for Function A, we found a controversial result that those who graduated from first tier schools and exhibited good performance were not likely to stay long and their resignation rate within one year was higher than those who graduated from other schools. Indeed, it makes sense to human resource staffs since they found the jobs in Function A including equipment engineering were not so attractive to the first tier graduates, though they could perform well. Thus, this company has adapted this finding and developed a job rotation strategy to allow the good performed engineers in Function A to be able to rotate to other functions before they quit the jobs.

(3) Work experience:

Employees with one or more years of work experience exhibited better performance as compared to those without experience. As recognized earlier, the employees with a master or higher degree showed better performances than the others and they were even better if had more work experience. Furthermore, the employees who majored in chemical or material engineering and had work experience would exhibit better performance.

On the other hand, the employees with one or more years of work experience had a higher resignation rate within three months as compared to those without experience, especially in Function C. Since those who had previous work experience would tend to compare their new jobs with previous ones, some of them may thus quit within first three months of probation and even return to their previous company. Thus, this company has adapted this finding to demand line manager and HR staff to pay attention to facilitate those who had previous work experience and also redesign some jobs in Function C.

Although we have derived several rules associated with turnover reasons, the HR staff thought these rules to be divergent for implementation after examining them. Nevertheless, future research will be done to redesign the survey for collecting true cause of turnover and thus conduct in depth analysis to improve retention rate.

4.7. Using discovered knowledge

Based on the findings and the interpretations through data mining and discussions, we developed specific recruiting strategies in order to fulfill the “right fit from the best” policy. Firstly, the company should recruit the students from the first tier schools through promoting their University Relationship Program (URP). Now this company has established this program with the four first tier universities in Taiwan and also extended it to major universities in USA including UC Berkeley, MIT, and Stanford. Secondly, this company has promoted a campaign for employee referral via cash award as well as professor referral system through the URP. Thirdly, some job functions are redesigned and their roles and responsibilities are adjusted to be competitive for attracting high-talents. Fourthly, job rotation mechanism is developed for cross functions to save high-performance talents from tedious jobs. Fifthly, mentoring system is developed for new hired employee and potential employees in the URP from the first tier schools.

5. Concluding remarks

High-tech companies rely on human capital to maintain competitive advantages. This study developed a data mining framework to extract useful rules from the relationships between personnel profile data and their work behaviors. Furthermore, we developed useful strategies with domain experts in the case company and most of the suggestions have been implemented. With an effective personnel selection process, organizations can find the suitable talents at the first time to improve retention rate and generate better performance. In addition, the mined results have also assisted in improving human resource management activities including job redesign, job rotation, mentoring, and career path development.

This study used applicants’ demographical data such as age, gender, marital status, education background, and work experience to predict their work performance and retention. Indeed, other demographical data and test results may be considered to improve the accuracy of the prediction or generate other potentially useful rules. Future study can be done to collect possible input variables such as address, the rank or scores in school, and number of owned licenses and to uncover buried relationships. Furthermore, although we have examined turnover reasons, the HR staff thought the derived rules to be divergent for implementation. Future research should be done to redesign the survey for collecting true turnover reasons for in depth analysis to help managers understand the root causes and thus take actions to effectively improve retention rate.

Decision tree is used for data mining in this study because it is easier to understand and it offers an acceptable level of accuracy. The empirical study has shown practical viability of this approach for extracting useful rules for human resource management in the semiconductor industry. Alternative data mining techniques such as neural network can be studied in future research to compare various approaches and may thus integrate them for better exploration of complex interrelationships among the input personnel variables and target work behaviors. Furthermore, this methodology can also be applied to other jobs like operators or management level jobs, and to other industries to find matched talents to enhance human capital. The validated results can be integrated into the human resource information system (HRIS) as a preliminary screening mechanism for the large amount of resumes gathered from external recruiting channels such as internet to reduce the workload of recruiters and save on both visible and invisible costs.

Acknowledgements

This research was supported by National Science Council, Taiwan (NSC94-2213-E-007-050) and Taiwan Semiconductor Manufacturing Company.

References

Appleyard, M. M., & Brown, C. (2001). Employment practices and semiconductor manufacturing performance. *Industrial Relations, 40*(3), 436–471.

Beckers, A. M., & Bsat, M. Z. (2002). A DSS classification model for research in Human Resource Information Systems. *Information Systems Management, 19*(3), 41–50.

Berry, M. J., & Linoff, G. (1997). *Data mining techniques: For marketing, sales, and customer support*. John Wiley & Sons.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48*, 299–337.

Braha, D., & Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor industry. *IEEE Transactions on Semiconductor Manufacturing, 15*(1), 91–101.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, P. J. (1984). *Classification and regression trees*. CA: Wadsworth International Group.

Chen, M. S., Han, J., & Yu, P. S. (1996). Data mining: An overview from a database perspective. *IEEE Transactions on Knowledge and Data Engineering, 8*(6), 866–883.

Chien, C. F., Chen, S., & Lin, Y. (2002). Using Bayesian network for fault location on distribution feeder of electrical power delivery systems. *IEEE Transactions on Power Delivery, 17*(13), 785–793.

Chien, C. F., Hsiao, A., & Wang, I. (2004). Constructing semiconductor manufacturing performance indexes and applying data mining for manufacturing data analysis. *Journal of the Chinese Institute of Industrial Engineers, 21*(4), 313–327.

Chien, C. F., Wang, I., & Chen, L. F. (2005). Using data mining to improve the quality of human resource management of operators in semiconductor manufactures. *Journal of Quality, 12*(1), 9–28.

Chien, C. F., Wang, W. C., & Cheng, J. C. (2007). Data mining for yield enhancement in semiconductor manufacturing and an empirical study. *Expert Systems with Applications, 33*(1), 1–7.

Chien, C. F., & Wu, J. (2003). Analyzing repair decisions in the site imbalance problem of semiconductor test machines. *IEEE Transactions on Semiconductor Manufacturing, 16*(4), 704–711.

Cho, V., & Ngai, E. (2003). Data mining for selection of insurance sales agents. *Expert Systems, 20*(3), 123–132.

Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39*, 27–34.

Fu, Y. (1997). Data mining: tasks, techniques and applications. *IEEE Potentials, 16*(4), 18–20.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. San Francisco, CA: Morgan Kaufman.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: John Wiley & Sons.

Hooper, R. S., Galvin, T. P., Kilmer, R. A., & Liebowitz, J. (1998). Use of an expert system in a personnel selection process. *Expert Systems with Applications, 14*(4), 425–432.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future – remembering the past. *Annual Review of Psychology, 51*, 631–664.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Kovach, K. A., & Cathcart, C. E. (1999). Human Resource Information Systems (HRIS): Providing business with rapid data access, information exchange and strategic advantage. *Public Personnel Management, 28*(2), 275–282.

Kusiak, A. (2001). A data mining tool for semiconductor manufacturing. *IEEE Transactions on Electronics Packaging Manufacturing, 24*(1), 44–50.

Liao, S. H. (2003). Knowledge management technologies and applications – literature review from 1995 to 2002. *Expert Systems with Applications, 25*, 155–164.

Lievens, F., Van Dam, K., & Anderson, N. (2002). Recent trends and challenges in personnel selection. *Personnel Review, 31*(5–6), 580–601.

Nussbaum, M., Singer, M., Rosas, R., Castillo, M., Flies, E., Lara, R., et al. (1999). Decision support system for conflict diagnosis in personnel selection. *Information & Management, 36*(1), 55–62.

Peng, C., & Chien, C. F. (2003). Data value development to enhance yield and maintain competitive advantage for semiconductor manufacturing. *International Journal of Service Technology and Management, 4*(4–6), 365–383.

Peng, J., Chien, C. F., & Tseng, B. (2004). Rough set theory for data mining for fault diagnosis on distribution feeder. *IEE Proceedings-Generation, Transmission, and Distributions, 151*(6), 689–697.

Pyle, D. (1999). *Data preparation for data mining*. San Francisco, CA: Morgan Kaufrnann.

Quinlan, J. R. (1986). Induction of decision tree. *Machine Learning, 1*(1), 81–106.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. San Mateo, CA: Morgan Kaufman.

Robertson, I. T., & Smith, M. (2001). Personnel selection. *Journal of Occupational and Organizational Psychology, 74*(4), 441–472.

Sattler, L., & Sohoni, V. (1999). Participative management: An empirical study of semiconductor manufacturing industry. *IEEE Transactions on Engineering Management, 46*(4), 387–398.

Shaw, M. J., Subramaniam, C., Tan, G. W., & Welge, M. E. (2001). Knowledge management and data mining for marketing. *Decision Support Systems, 31*(1), 127–137.

Shiue, Y. R., & Su, C. T. (2003). An enhanced knowledge representation for decision tree based learning adaptive scheduling. *International Journal of Computer Integrated Manufacturing, 16*(1), 48–60.

Wei, C., & Chiu, I. (2002). Turning telecommunications call details to churn prediction: a data mining approach. *Expert Systems with Applications, 23*(2), 103–112.

Wu, C. H., Kao, S. C., Su, Y. Y., & Wu, C. C. (2005). Targeting customers via discovery knowledge for the insurance industry. *Expert Systems with Applications, 29*(2), 291–299.