2009 Ninth International Conference on Intelligent Systems Design and Applications

Potential Data Mining Classification Techniques for Academic Talent Forecasting

Hamidah Jantan
Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM) Terengganu,
23000 Dungun, Terengganu, Malaysia
e-mail: hamidahjtn@tganu.uitm.edu.my

Abdul Razak Hamdan and Zulaiha Ali Othman
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor, Malaysia
e-mail: {arh,zao}@ftsm.ukm.my

***Abstract*--Classification and prediction are among the major techniques in Data mining and widely used in various fields. In this article we present a study on how some talent management problems can be solved using classification and prediction techniques in Data mining. By using this approach, the talent performance can be predicted by using past experience knowledge discovered from the existing database. In the experimental phase, we have used selected classification and prediction techniques to propose the appropriate techniques from our training dataset. An example is used to demonstrate the feasibility of the suggested classification techniques using academician performance data. Thus, by using the experiments results, we suggest the potential classification techniques for academic talent forecasting.**

Keyword-Data Mining, Classification Techniques, Academic Talent, and Forecasting.

I. INTRODUCTION

Data mining is a step in Knowledge Discovery in Database (KDD) approach that is used to extract and discover meaningful knowledge from large amount of data. Besides that, among the major functions of Data mining are classification and prediction; concept description; association; cluster analysis; outlier analysis; trend and evaluation analysis; statistical analysis and many others. Classification and prediction are among the popular function in Data mining. The technique is supervised learning, which is class level or prediction target is known. There are many areas that have adapted this approach such as finance, medical, marketing, stock, telecommunication, manufacturing, health care, customer relationship and etc. However, the application of Data mining has not attracted much attention in Human Resource (HR) field [3, 4]. HR data can provide a rich resource for knowledge discovery and for decision support tool development. Nowadays, an organization has to struggle effectively in term of cost, quality, service or innovation. All these depend on having enough right people with the right skills, employed in the appropriate locations at appropriate point of time that is known as talent management.

Recently, among the challenges of HR professionals are managing an organization talent. This task involves a lot of managerial decisions. Sometime, these types of decisions are very uncertain and difficult; and depend on various factors like human experience, knowledge, preference and judgment. Besides that, the process to identify the existing talent in an organization is among the top talent management challenges and serious issue[5]. Talent management is defined as an outcome to ensure the right person is in the right job[6]. Employees in an organization are evaluated based on the position that he/she holds, and the post is represented by the talent ability that he/she has. For that reason, this study aims to use classification techniques when the class level (talent position) is known in order to handle this issue. In this study, we attempt to use employees in higher education institution as our dataset especially from academician data. Therefore, the purpose of this paper is to suggest possible classification techniques for talent forecasting through some experiments using the selected classifier algorithms.

This paper is organized as follows. The second section describes the related work on Data mining in HR especially for talent management, classification and prediction in Data mining; and the possible methods for classification. The third section discusses the experiment setup in this study. Section 4 shows some experiments results and analysis. Finally, the paper ends at Section 5 with the concluding remarks and future research directions are suggested.

II. RELATED WORK

A. Data Mining in Human Resource

Recently, there are some researches that show great interest on solving HRM problems using Data mining approach[3]. Table I lists some of the applications in human resource that use Data mining techniques, and it shows that there are few discussions or researches on classification and prediction in human resource domain. Moreover, Data mining technique is usually used in personnel selection to choose the right candidates for a job. Classification and prediction applications in HRM are infrequent and there are

978-0-7695-3872-3/09 $26.00 © 2009 IEEE
DOI 10.1109/ISDA.2009.64

some examples such as to predict the length of service, sales premium, to persistence indices of insurance agents and analyze miss-operation behaviors of operators[4]. Due to these reasons, this study attempts to use Data mining techniques to forecast potential employees as a part of talent management task.

TABLE I DATA MINING IN HR APPLICATIONS

Data Mining method	Activity in HRM
Fuzzy Data Mining and Fuzzy Artificial Neural Network	Employee development – Project Assignment [7]
Decision tree	Personnel selection [4], Job attitudes [8]
Association rule mining	Employee Development – Training [9]
Rough Set Theory	Personnel Selection – Recruit and Retain Talents [10]
Fuzzy Data Mining	Personnel Selection [11]

Data mining tasks are generally categorized as clustering, association, classification and prediction[3, 4]. Over the years, data mining has evolved various techniques to perform the tasks that include database oriented techniques, statistic, machine learning, pattern recognition, neural network, rough set and etc. Data mining technique has been applied in many fields, but its application in Human Resource Management(HRM) is very rare[12]. Recently, there are some researches that show interest in solving HRM problems using data mining approach[1, 3]. However, until now there have been few discussions on talent management such as talent forecasting, project assignment and talent recruitment using Data mining approach. Due to these reasons, this study attempts to use Data mining techniques i.e. classification and prediction in order to identify the potential talent by predicting the talent based on past experience knowledge.

B. Talent Management and Data Mining

In any organization, talent management has become an increasingly crucial approach in HR functions. Talent is considered as the capability of any individual to make a significant difference to the current and future performance of the organization [13]. In fact, managing talent involves human resource planning that emphasizes processes for managing people in organization. Besides that, talent management can be defined as a process to ensure leadership continuity in key positions and encourages individual advancement; and decision to manage supply, demand and flow of talent through human capital engine [6].

Talent management is very crucial and needs some attention from HR professionals. TP Track Research Report has found that among the top current and future talent management challenges are developing existing talent; forecasting talent needs; attracting and retaining the right leadership talent; engaging talent; identifying existing talent; attracting and retaining the right leadership and key contributor; deploying existing talent; lack of leadership capability at senior levels and ensuring a diverse talent pool [5]. This study focuses on one of the talent management challenge that is to identify the existing talent regarding the key talent in an organization by predicting their performance. In this case, we use the past data from the employee database to implement the classification and prediction process. The talent management process consists of recognizing the key talent areas in the organization, identifying the people in the organization who constitute key talent, and conducting development activities for the talent pool to retain and engage them and also have them ready to move into more significant roles [6] as illustrated in Fig.1. These processes involve HR activities that need to be integrated into an effective system [14].

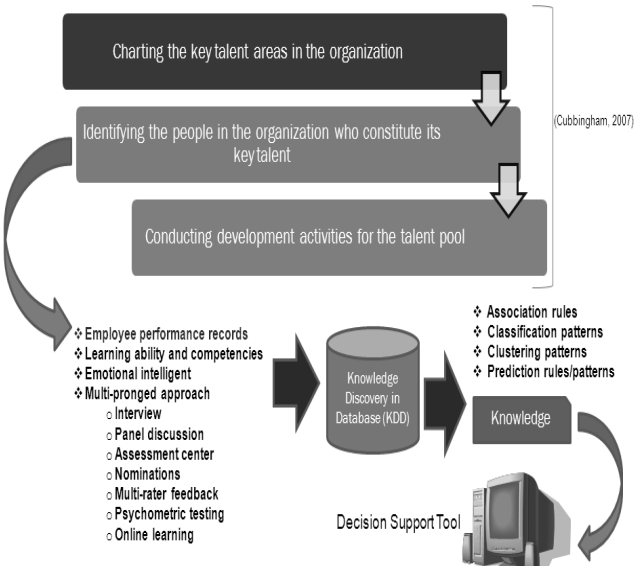

Figure 1. Data Mining for Talent Management [1]

Recently, with the new demand and increased visibility, HRM seeks a more strategic role by turning to Data mining methods [3]. This can be done by identifying generated patterns from the existing data in HR databases as useful knowledge. Thus, this study concentrates on identifying the patterns that relate to the talent. The patterns can be generated by using some of the major data mining techniques such as clustering to list the employees with similar characteristics, to group the performances and etc. From the association technique, patterns that are discovered can be used to associate the employee's profile for the most appropriate program/job, associated with employee's attitude to performance and etc. In prediction and classification, the pattern can be used to predict the percentage accuracy in employee's performance, behavior, and attitudes, predict the performance progress throughout

the performance period, and also identify the best profile for different employee and etc. [15]. The matching of data mining problems and talent management needs are very crucial. Therefore, it is very important to determine the suitable data mining techniques (Fig. 2).

C. Classification and Prediction

Database or data warehouse are rich with hidden information that can be used to provide intelligent decision making. Intelligent decision refers to the ability to make automated decision that is quite similar to human decision. Prediction and classification abilities are among the methods that can produce intelligent decision. Currently, many classification and prediction methods have been proposed by researchers in machine learning, pattern recognition, and statistics. This study focuses our discussion on classification and prediction methods in machine learning. Prediction and classification in Data mining are two forms of data analysis that can be used to extract models to describe important data classes or to predict future data trends[16].

The classification process has two phases; the first phase is learning process whereby training data are analyzed by classification algorithm. Learned model or classifier is represented in the form of classification rules. The second phase is classification, and test data are used to estimate the accuracy of classification rules. If the accuracy is considered acceptable, the rules can be applied to the classification of new data (Fig. 3). Some of the techniques that are used for data classification are decision tree, Bayesian methods, Bayesian network, rule-based algorithms, neural network, support vector machine, association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets, fuzzy logic. In this study, our discussion focuses on three classification techniques i.e. decision tree, neural network and k-nearest-neighbor. However, decision tree and neural network are found useful in developing predictive models in many fields[17].

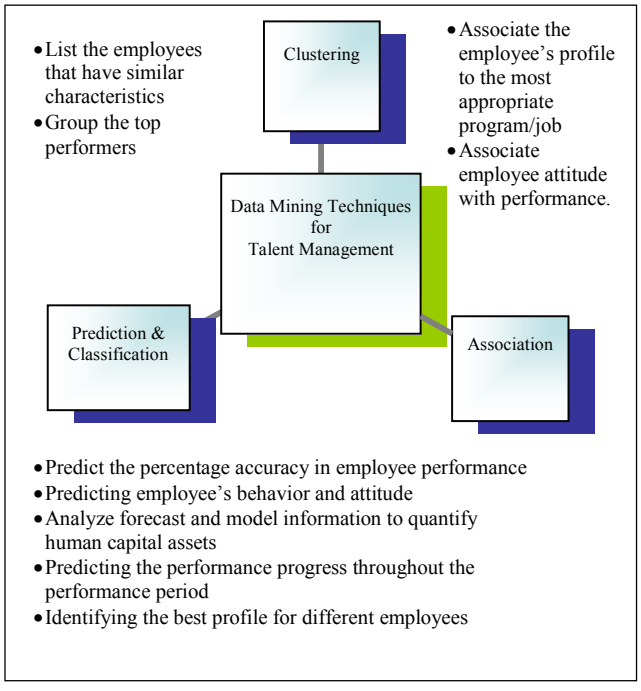

Figure 2. Data Mining Techniques for Talent Management (Enhancement from Ranjan [2])

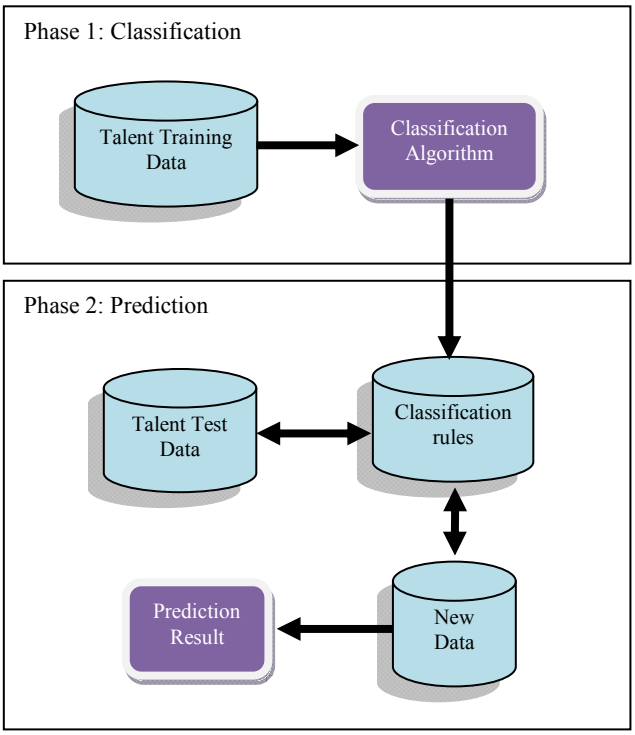

Figure 3. Classification and Prediction Process

The advantage of decision tree technique is that it does not require any domain knowledge or parameter setting, and is appropriate for exploratory knowledge discovery. The second technique is neural-network which has high tolerance of noisy data as well as the ability to classify pattern on which they have not been trained. It can be used when we have little knowledge of the relationship between attributes and classes. Next, the K-nearest-neighbor technique is an instance-based learning using distance metric to measure the similarity of instances. All these three classification techniques have their own advantages and for that reasons, we attempt to explore these classification techniques for HR data. Table II summarizes the potential techniques such as decision tree, neural network and nearest neighbor. In this study, we attempt to use C4.5 and Random Forest for decision tree category; Multilayer Perceptron (MLP) and Radial Basic Function Network (RBFC) for neural network category; and K-Star for the nearest neighbor category.

TABLE II. POTENTIAL DATA MINING CLASSIFICATION TECHNIQUES

Data Mining Techniques	Classification Algorithm
Decision Tree	• ***C4.5*** (Decision tree induction – the target is nominal and the inputs may be nominal or interval. Sometimes the size of the induced trees is significantly reduced when a different pruning strategy is adopted). • ***Random forest*** (Choose a test based on a given number of random features at each node, performing no pruning. Random forest constructs random forest by bagging ensembles of random trees).
Neural Network	• ***Multi Layer Perceptron*** (An accurate predictor for underlying classification problem. Given a fixed network structure, we must determine appropriate weights for the connections in the network). • ***Radial Basic Function Network*** (Another popular type of feed forward network, which has two layers, not counting the input layer, and differs from a multilayer perceptron in the way that the hidden units perform computations).
Nearest Neighbor	• ***K*Star*** **(**An instance-based learning using distance metric to measure the similarity of instances and generalized distance function based on transformation

III. EXPERIMENT SETUP

This experiment, attempts to identify the talent patterns in the existing HR databases and several classification techniques are used for the simulated employee data. The selected classification techniques are based on the common techniques used for classification and prediction especially in Data mining. As mentioned earlier, the classification techniques chosen are neural network which is quite popular in data mining community and used as pattern classification technique[18]. The decision tree as 'divide-and–conquer' approach from a set of independent instances for classification and the nearest neighbor for classification that are based on the distance metric are summarized in Table II. The process of classification includes the input variables i.e. talent factors for academic staff; and the outcome of the classification process i.e. talent patterns are shown in Figure 4. In this study, the performance factors for the case study are based on academician performance. For that reason, the most important performance factors are extracted from the previous performance, knowledge and expertise records. Besides the performance factors, the background and management skill are also important to identify the possible talent for that job. In this experiment, the training dataset contains 53 related attributes from five performance factors that are demonstrated in Table III.

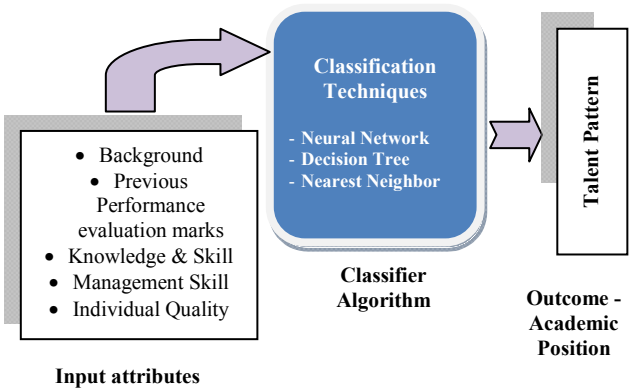

Figure 4. Classification for Academic Talent

TABLE III. ATTRIBUTES FOR ACADEMIC TALENT

Factor and Attributes	Variable Name	Meaning
Background (7)	D1,D2,D3,D5,D6, D7,D8	Age ,Race, Gender, Year of service, Year of Promotion 1, Year of Promotion 2, Year of Promotion 3
Previous performance evaluation (15)	DP1,DP2,DP3, DP4,DP5,DP6, DP7,DP8,PP9, DP10, DP11,DP12, DP13,DP14, DP15	Performance evaluation marks for 15 years
Knowledge and skill (20)	PQA,PQC1,PQC2, PQC3,PQD1, PQD2,PQD3,PQE1, PQE2,PQE, PQE4,PQE5,PQF1, PQF2,PQG1, PQG2,PQH1,PQH2, PQH3,PQH4	Professional qualification (Teaching, supervising, research, publication and conferences)
Management skill (6)	PQB,AC1,AC2,AC3, AC4,AC5	Student obligation and administrative tasks
Individual Quality (5)	T1,T2,SO,AA1,AA2	Training, award and appreciation

However, due to confidentiality and security of data, for the exploratory purposes, we simulate one hundred performance data that are based on the talent performance factors. The classification technique used is based on 10 fold cross validation training and test dataset. In this experiment, the Data mining tools used are WEKA and

ROSETTA toolkit. This experiment has two phases; the first phase is to identify the possible techniques using selected classifier algorithm for full attributes of data. In this case, we use all attributes that we defined before for full dataset. In this experiment, we concentrate our study on the accuracy of the classifier to identify suitable classifier algorithm for the sample dataset.

The accuracy of classifier is based on the percentage of test set samples that are correctly classified. The second phase of experiment is to compare the accuracy of classifier for attribute reduction. In this case, Boolean reasoning method is used to select the relevant attributes. Attribute reduction phase is divided into two stages. The first stage is attribute reduction for shortest length attribute which used by many researches, in order to select the important attribute for the data set. The second stage is the process for all important attribute from attribute reduction process which known as combination attributes. In this case, we attempt to study the accuracy of the classifier for all important attributes.

IV. RESULTS AND DISCUSSION

In this experiment, the accuracy of classification techniques is based on the selected classifier algorithm. The accuracy for each of the classifier algorithm for full attributes using sample dataset shown in Table IV. The results for full attributes show the highest accuracy are C4.5 (95.1%) and K-Star (92.1%) and can be considered as indicators for the suitable classification techniques.

TABLE IV. THE ACCURACY FOR FULL ATTRIBUTES

Classifier Algorithm	*Accuracy*
C4.5	95.14%
Random forest	74.91%
Multi Layer Perceptron (MLP)	87.16%
Radial Basis Function Network	91.45%
K-Star	92.06%

The second phase of the experiment is considered as a relevant analysis process in order to find out the accuracy of the selected classification technique using sample dataset with attribute reduction. The purpose of attribute reduction is to choose the most relevant attribute only in the dataset. The reduction process is implemented using Boolean reasoning technique. With attribute reduction, we can decrease the preprocessing and processing time and space. Table V shows the relevant analysis results for attribute reduction, five (5) attributes are selected which are from the background input. By using these attributes reduction variables, the second phase of experiment is implemented. The purpose of this experiment is to find out the accuracy of the classification techniques with attribute reduction using the shortest length attributes and combination of the important attribute from reduction process.

TABLE V. REDUCTION ATTRIBUTES

Variable Name	**Meaning**
D1,D5,D6,D7,D8	Age, Year of service, Year of Promotion 1, Year of Promotion 2, Year of Promotion 3

Table VI shows the accuracy of the classification technique with attribute reduction for both methods. The C4.5 classifier algorithm is among the highest accuracy for each of the attribute reduction experiments. In addition, the C4.5 for decision tree and K-star for the Nearest Neighbor have the highest percentage of accuracy in the first experiment (Table IV) i.e. for full attributes but the accuracy has declined in attribute reduction experiment. The percentage of accuracy for the shortest length and combination attribute experiment shown in Table VI. The result indicates more attributes used and will affect the accuracy of the classifier. Consequently, this result also shows, may be most of the attributes in dataset are important in these experiments. However, with the combination of attributes from reduction process, the accuracy of classifier is a little bit higher compared to the shortest length attributes.

TABLE VI. THE ACCURACY FOR ATTRIBUTE REDUCTION

Classifier Algorithm	*Shortest length*	*Combination of Attributes*
C4.5	61.06%	95.63%
Random forest	58.85%	86.50%
Multi Layer Perceptron (MLP)	55.32%	79.49%
Radial Basis Function Network	59.52%	84.41%
K-Star	60.22%	78.40%

For future work, these experiment results should be tested using relevancy analysis process to validate these findings whether the number of attributes will affect the accuracy of the classifier. In addition, from these experiments, we observe the great potential to use C4.5 classification technique in the next stage of Data mining process which is known as prediction. This result also shows the suitability of the classification techniques or classifier for the dataset (Fig. 5). For these reasons, in the next experiment setup other decision tree techniques should be conducted to support this finding.

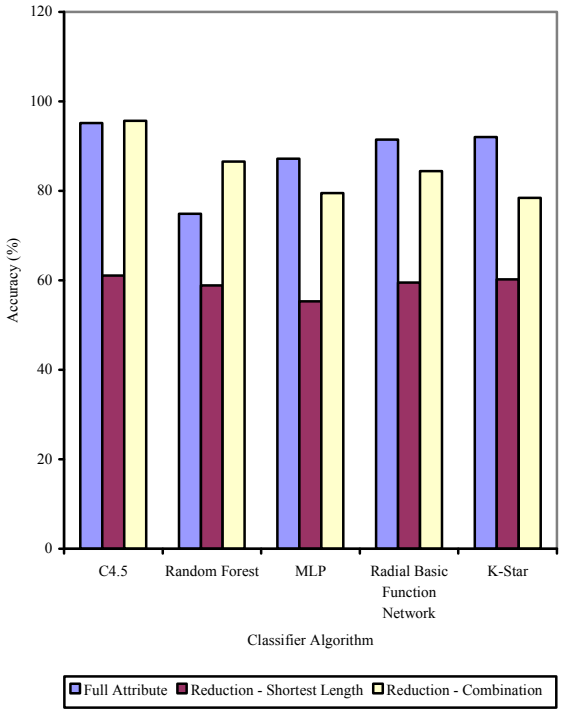

Figure 5. The Accuracy of Classifier Algorithm for Full Attributes and Attribute Reduction

V. CONCLUSION AND FUTURE WORK

This article has described the significance of the study using Data mining for talent management especially for classification and prediction. However, there should be more Data mining techniques applied to the different problem domains in HR field of research in to broaden our horizon of academic and practice work on Data mining in HR. Other Data mining techniques such as Support Vector Machine (SVM), Fuzzy logic and Artificial Immune System (AIS) should also be considered for future work on classification techniques using the same dataset.

In some cases, the attribute relevancy also reacts as a factor on the accuracy of the classifier algorithm. In the next experiments, the attribute reduction process should be implemented using other reduction techniques to confirm these findings. C4.5 classifier has the highest accuracy in the experiment; the accuracy of other decision tree classifier should also be tested in order to validate these findings. In addition, C4.5 classifier algorithm is the potential classifier in this experiment. Thus, this technique should be applied in the next prediction phase to construct classification rules. These generated classification rules can be used to predict the potential academic talent. In conclusion, the ability to continuously change and obtain new understanding of the classification and prediction in HR researches has thus, become the major contribution to HR Data mining.

ACKNOWLEDGEMENT

This research was conducted as a part of the eScience project funded by MOSTI (Ministry of Science, Technology and Innovation), Malaysia (01-01-01-SF0236).

REFERENCES

[1] J. Hamidah, H. Abdul Razak, and A. O. Zulaiha, "Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application," in World Academy of Science, Engineering and Technology, Penang, Malaysia, 2009.

[2] J. Ranjan and K. Malik, "Effective educational process: a data mining approach," VINE: The Journal of Information and Knowledge Management Systems, vol. 37, pp. 502-515, 2007.

[3] J. Ranjan, "Data Mining Techniques for better decisions in Human Resource Management Systems," International Journal of Business Information Systems, vol. 3, pp. 464-481, 2008.

[4] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Systems and Applications, vol. 34, pp. 380-290, 2008.

[5] A TP Track Research Report "Talent Management: A State of the Art," Tower Perrin HR Services 2005.

[6] I. Cubbingham, "Talent Management: Making it real," Development and Learning in Organizations, vol. 21, pp. 4-6, 2007.

[7] M. J. Huang, Y. L. Tsou, and S. C. Lee, "Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge," Knowledge-Based Systems, vol. 19, pp. 396-403, 2006.

[8] K. Y. Tung, I. C. Huang, S. L. Chen, and C. T. Shih, "Mining the Generation Xer's job attitudes by artificial neural network and decision tree - empirical evidence in Taiwan," Expert Systems and Applications, vol. 29, pp. 783-794, 2005.

[9] K. K. Chen, M. Y. Chen, H. J. Wu, and Y. L. Lee, "Constructing a Web-based Employee Training Expert System with Data Mining Approach," Paper in The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007), 2007.

[10] C. F. Chien and L. F. Chen, "Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing," IEEE Transactions on Semiconductor Manufacturing, vol. 20, pp. 528-541, 2007.

[11] W. S. Tai and C. C. Hsu, "A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method," 2005.

[12] C. F. Chien and L. F. Chen, "Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry," Expert Systems and Applications, vol. 34, pp. 380-290, 2008.

[13] M. Lynne, "Talent Management Value Imperatives: Strategies for Execution," The Conference Board 2005.

[14] CHINA UPDATE, "HR News for Your Organization: The Tower Perrin Asia Talent Management Study," 2007.

[15] J. Hamidah, H. A. Razak, and A. O. Zulaiha, "Data Mining Techniques for Performance Prediction in Human Resource Application," in 1st Seminar on Data Mining and Optimization, Bangi, Selangor, 2008.

[16] J. Han and M. Kamber, Data Mining : Concepts and Techniques. San Francisco: Morgan Kaufmann Publisher, 2006.

[17] G. K. F. Tso and K. K. W. Yau, "Predicting electricity energy consumption: A comparison of regression analysis, decision tree and neural networks," Energy, vol. 32, pp. 1761-1768, 2007.

[18] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques. San Francisco: Morgan Kaufmann Publishers, 2005.