Applying Data Mining Classification Techniques for Employee's Performance Prediction

Hamidah Jantan[1], Mazidah Puteh[1], Abdul Razak Hamdan[2] and Zulaiha Ali Othman[2]

[1]Faculty of Computer and Mathematical Sciences
Universiti Teknologi MARA (UiTM) Terengganu,
23000 Dungun, Terengganu, MALAYSIA
{hamidahjtn, mazidahputeh}@tganu.uitm.edu.my

[2]Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia (UKM)
43600 Bangi, Selangor, MALAYSIA
{arh,zao}@ftsm.ukm.my

ABSTRACT

The valuable knowledge can be discovered through data mining process. In data mining, classification is one of the major tasks to impart knowledge from huge amount of data. This technique is widely used in various fields, but it has not attracted much attention in Human Resource Management (HRM). This article presents a study on the implementation of data mining approach for employee development regarding to their future performance. By using this approach, the performance patterns can be discovered from the existing database and will be used for future performance prediction in their career development. In the experimental phase, we have used selected classification techniques to propose the appropriate technique for the dataset. An experiment is carried out to demonstrate the feasibility of the suggested classification techniques using employee's performance data. Thus, the experiment results, we suggest the potential classification techniques and the possible prediction model for employee's performance forecasting.

Keywords
Data Mining, Classification Techniques, Employee's Performance, Prediction.

1.0 INTRODUCTION

Nowadays, in the K-Era, knowledge is a valuable asset and among the crucial issues to address. Knowledge can be discovered through many approaches and one of them is by using data mining technique. In data mining, tasks such as classification, clustering and association are used to discover implicit knowledge from huge amount of data. Classification technique is a supervised learning technique in machine learning, which the class level or the target is already known. There are many fields adapted this approach as their problem solver method, such as finance, medical, marketing, stock market, telecommunication, manufacturing, health care, customer relationship, education and some others. Nevertheless, the application of data mining has not attracted much attention in Human Resource Management (HRM) field (Chien & Chen, 2008; Ranjan, 2008). The vast amount of data in HRM can provide a rich resource for knowledge discovery and for decision support system development. Besides that, the valuable knowledge discovered from data mining process should be considered as part of knowledge management issues. In any organization, they have to struggle effectively in term of cost, quality, service or innovation. The success of these tasks depends on having enough right people with the right skills, employed in the appropriate locations at appropriate point of time. This is categorized as part of the talent management task in HRM. In addition, talent management is a process to ensure the right person is in the right job (Cubbingham, 2007).

Recently, among the challenges of human resource professionals are managing an organization talent which involves a lot of managerial decisions. These types of decision are very uncertain and difficult. It depends on various factors like human experiences, knowledge, preferences and judgments. Besides that, the process to identify the existing talent in an organization is among the top talent management issues and challenges (A TP Track Research Report 2005). Employees in an organization are evaluated based on their performance in order to represent their talent ability. For that reason, this study aims to use classification techniques to classify the employee's performance. In this case, the class level for the performance is whether the employee gets recommendation for promotion or not. In this study, we use employee's performance data from selected organization as our dataset. Therefore, the purpose of this paper is to suggest the possible classification techniques for employee future performance through some experiments using the selected classification

algorithms. As a result, by using proposed classifier, we generate prediction model which can be used for employee's performance prediction.

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

2.0 RELATED WORK

2.1 Data Mining in Human Resource

Data mining tasks are generally categorized as clustering, association, classification and prediction(Chien & Chen, 2008; Ranjan, 2008). 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 theory, support vector machine (SVM), artificial immune system (AIS) and some other techniques. Data mining technique has been applied in many fields, but its application in Human Resource Management(HRM) is very rare(Chien & Chen, 2008). Recently, there are some researches that show interest in solving HRM problems using Data mining approach (Jantan, Hamdan, & Othman, 2009; Ranjan, 2008). Table I lists some of the applications in HR that use data mining tasks, and it shows that there are few discussions on that. Moreover, data mining technique is usually used in personnel selection, in order to choose the right candidates for a job. The use of data mining techniques in HRM are infrequent and there are some examples such as to predict the length of service, sales premium, persistence indices of insurance agents and to analyze miss-operation behaviors of operators(Chien & Chen, 2008).

Besides that, there are very few discussions or research on the uses of data mining in talent management such as for talent forecasting, project assignment and talent recruitment. Due to these reasons, this study attempts to use classification techniques in data mining to determine the employee's performance by predicting their performance based on the past experience knowledge from employee databases.

Table I: Data Mining Task in HR Applications

Data Mining Task	Activity in HRM
Classification	Personnel selection (Chien & Chen, 2008),
	Job attitudes (Tung, Huang, Chen, & Shih, 2005)
	Personnel Selection – Recruit and Retain Talents (Chien & Chen, 2007)
Association	Training (Chen, Chen, Wu, & Lee, 2007)
Classification and Prediction	Project Assignment (Huang, Tsou, & Lee, 2006)
Classification and Association	Personnel Selection (Tai & Hsu, 2005)

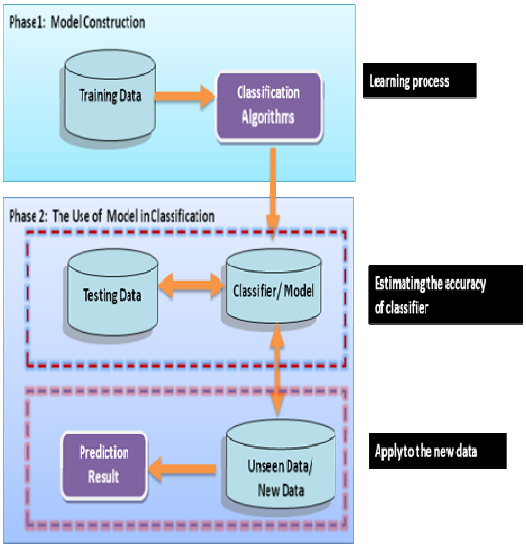
2.2 Classification in Data Mining

Databases or data warehouses 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. Classification and prediction are some of 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. In this study, we are focusing on classification methods in data mining as part of machine learning process.

Classification and prediction 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(Han & Kamber, 2006). The classification process has two phases; the first phase is learning process where the training data are analyzed by the classification algorithm. Learned model or classifier is represented in the form of classification rules. The second phase is classification process, where the test data are used to estimate the accuracy of classification model or classifier. If the accuracy is considered acceptable, the model can be applied to the new data to know the prediction result (Figure 1).

There are many techniques that can be used for classification such as 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 and fuzzy logic. In this study, our discussion focuses on the three main classification techniques i.e. decision tree, neural network and k-nearest-neighbor. Decision tree and neural network are found useful in developing predictive models in many fields(Tso & Yau, 2007).

Figure 1: Fundamental Data Mining Process for Classification and Prediction



The advantages of decision tree technique are such that it does not require any domain knowledge, parameter setting, and is appropriate for knowledge discovery. The second technique is neural-network which has high tolerance to 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 that uses 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 of decision tree, neural network and nearest neighbor algorithm. In this study, we use C4.5 and Random Forest for decision tree; Multilayer Perceptron (MLP) and Radial Basic Function Network (RBFC) for neural network; and K-Star for the nearest neighbor.

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.

3.0 EXPERIMENT SETUP

In the experimental phase, we attempt to impart employee's performance patterns in the existing HR databases using selected common classification techniques. In the model construction phase, the main classification techniques are neural network, decision tree and nearest neighbor. Neural network is quite popular in data mining community for pattern classification technique(Witten & Frank, 2005). The decision tree is a 'divide-and–conquer' approach from a set of independent instances and the nearest neighbor is based on the distance metric (Table II). The process of classification uses input variables i.e. performance factors; and the outcome is the employee's performance pattern that shows the status of promotion. The attributes for dataset in this experiment as shown in Table III.

Table III: Attributes for Performance Prediction

Attribute	Description
Category	P – Professional, S - Support Staff
Gender	Male and Female
Qualification	Doctorate, Master, Bachelor, Diploma and Certificate
$PK_{1...6}$	Work Outcome (50%)
$PM_{1...6}$	Knowledge and Skill (25%)
$KP_{1...6}$	Individual Quality (20%)
$KS_{1...6}$	Activities and Contribution (5%)
$YEAR_{1...6}$	Evaluation mark (100%)
Target/Class	Recommendation for promotion (Yes or No)

The attributes for training dataset are selected based on the related factors for employee performance (ExecutiveBrief, 2008) as illustrated in Figure 2. These attributes are extracted from the individual factors component such that work outcome; knowledge and skill; individual quality; and activities and contribution. In this study, the performance factors are extracted from standard performance used for Malaysian public sector organizations. Besides the performance factors, some background information are also considered as part of parameters for the dataset. The aim of this experiment is to proposed the suitable classification algorithm for the dataset; and to construct employee's performance forecasting model using proposed classifier. The generated forecasting model contains classification rules for performance prediction. The classification rules will show us about the interesting or important attributes for the dataset. Besides that, the forecasting model will be used to determine whether the employee is recommended for promotion or not based on their performance.

In this experiment, the training dataset contains 33 related attributes from background information and performance factors which is demonstrated in Table III. The dataset contains 655 records from six years (2003-2008) performance evaluation marks. Each record holds evaluation marks for selected factors and the total mark for each of the year. The dataset is organized into 10 fold cross validation training and test dataset. 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 classification algorithm for full attributes of data. In this case, all attributes th defined in Table III are used for full dataset. We concentrate our study on the accuracy of the classifier in order to identify the suitable classification algorithm for the dataset. The accuracy of the classifier is based on the percentage of test set samples that are correctly classified. In this case, the classifier with the highest accuracy is considered as the most suitable classifier for the dataset. The second phase of the experiment is to construct employee's prediction model using the proposed classification algorithm. The generated classifier can be embedded into decision support system for employee's performance prediction which is based on the previous performance.

4.0 RESULT AND DISCUSSION

In this experiment, the accuracy of classification techniques is measured by averaging the accuracy from 10 fold cross validation datasets. As part of the classification process, the classifier generated by each classification technique must be applied to the unseen data. This process is known as the use of model phase, which shows the percentage of correctly classified instances or the accuracy of the models. The average accuracy obtained from each of the algorithms for the dataset is shown in Table IV. The result of this experiment showed that all of the classifiers have moderate accuracy, which is more than 70%. In many cases, the moderate accuracy is considered as an acceptable accuracy. In this experiment, the dataset produced acceptable models for each of selected classification algorithms. In order to choose the suitable classifier in data mining, the accuracy of the model is used to determine the most suitable classifier for the dataset. As shown in Table IV, The C4.5/J4.8 classifier has 79.49% accuracy, which is the highest accuracy among the selected classifiers. Due to those reasons, the C4.5/J4.8 classifier was selected as a classifier for the dataset.

Table IV: The Accuracy of Classifier

Classifier Algorithm	Accuracy(%)
C4.5 /J4.8	79.49
Random forest	72.21
Multi Layer Perceptron(MLP)	70.25
Radial Basis Function Network	70.53
K-Star	79.34

That C4.5/J4.8 comes from the family of decision tree classification technique. The rules generated from this classifier is easy to understand, human readable and has a very straightforward interpretation. This technique can produce rules in tree structures and rule-sets; and construct a tree for the purpose of improving the prediction accuracy (Becerra-Fernandez, Zanakis, & Walczak, 2002; Delen, Walker, & Kadam, 2005). Besides that, the C4.5/C5.0/J48 classifier is among the popular and powerful decision tree classifier (Becerra-Fernandez et al., 2002; Delen et al., 2005; Kumar & Ravi, 2007; Tso & Yau, 2007). By using C4.5/J4.8 classifier, a part of generated classifier or rules is shown in Figure 3. The leaves represent the classification results or the target class. In this study, the target class is the recommendation for promotion (Yes/No).

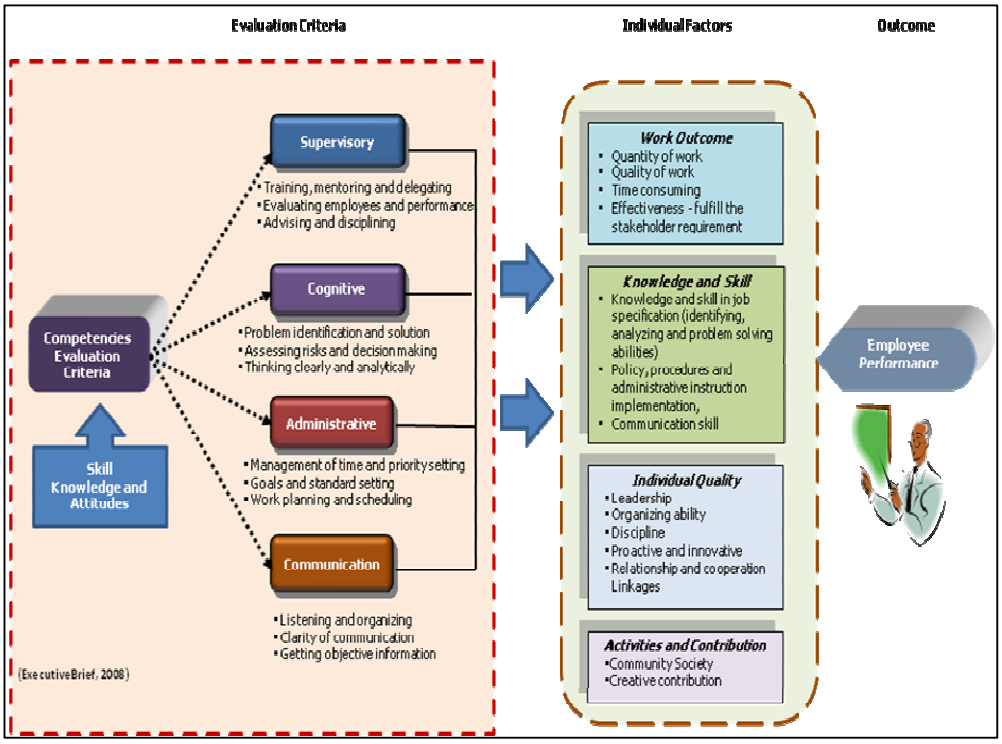


Figure 2: Related Factors for Employee Performance

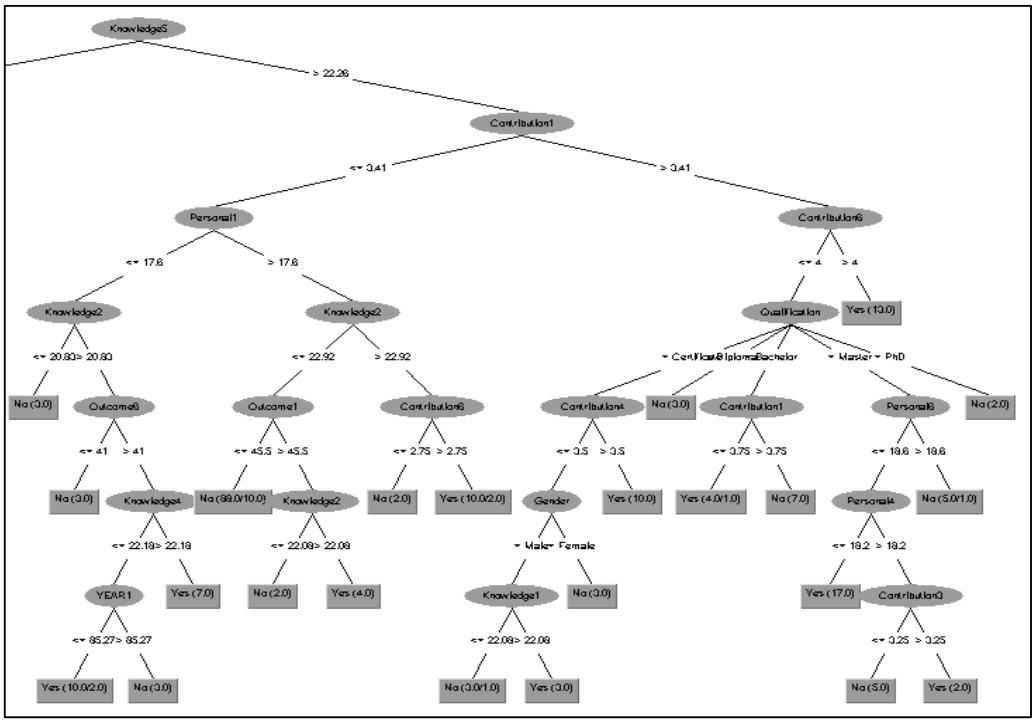


Figure 3: Sample of Rules using Decision Tree

The C4.5 classifier uses two heuristics criteria to rank possible tests. One is information gain by using attribute selection measure and the default gain ratio that divides information gain by the information provided by the test outcomes. For that reason, C4.5 classifier also can be used to determine the important or interesting attributes from the dataset. In this study, the important attributes are identified through the number of hits for each of attributes in the generated classification rules. The number of hits for each factors is shown in Table V. From 33 selected attributes, only 23 attributes are hits in the classification process and *Knowledge2* attribute has highest hit. Besides that, the knowledge and skill; and activities and contribution factors has top hit among the selected factors. This result indicates the vital of those factors for employee's performance, where all the attributes are hit and considered for predicting employee's performance.

In order to validate this finding, in future work, the experiments related to relevancy analysis should be extended to get the relevancy of the attributes. That experiment can be used to determine whether the attribute reduction would affect the performance of the classifier or not.

Table V: Interesting Attributes

Factors	Attributes Hit	Attribute Name
Category(1)	1	*Category(1)*
Gender(1)	1	*Gender(2)*
Qualification(1)	1	*Qualification(1)*
Work Outcome (6)	4 (Year1, Year2, Year4, Year6)	*Outcome1(2), Outcome2(2), Outcome4(2), Outcome6(2)*
Knowledge and Skill (6)	6 (Year1, Year2, Year3, Year4, Year5, Year6)	*Knowledge1(2), Knowledge2(4), Knowledge3(1), Knowledge4(2), Knowledge5(1), Knowledge6(2)*
Individual Quality (6)	3 (Year1, Year4, Year6)	*Personal1(2), Personal4(1), Personal6(1),*
Activities and Contribution (6)	6 (Year1,Year2, Year3,Year4, Year5, Year6)	*Contribution1(2), Contribution2(1), Contribution3(2), Contribution4(1), Contribution5(1), Contribution6(2)*
Year Evaluation mark (6)	1 (Year 1)	*YEAR1*

In this experiment, we observe the great potential of C4.5/J4.8 as classification technique for employee's performance prediction. In the next stage of data mining classification techniques analysis, it should be further analyzed with other related datasets. However, this result shows the suitability of the classification techniques for the selected dataset. The accuracy of other decision tree classification algorithms should also be experimented in order to validate these findings. For that reason, in future work, other decision tree techniques such as NBTree, SimpleCart, REPTree, BFTree and others will be tested to support this finding.

5.0 CONCLUSION

This article has described the significance of the study on data mining for employee's performance prediction. Nevertheless, there should be more data mining techniques applied to the different problem domains in HR field of research to broaden the horizon of academic and practice work on data mining in HR. Other data mining techniques such as Support Vector Machine (SVM), Fuzzy logic, Artificial Immune System (AIS) and many others 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 to weight the accuracy of the classification. In the next experiment, the attribute analysis should be conducted using other reduction techniques in order to strengthen these findings. As we can see, the C4.5/J4.8 classifier has highest accuracy in the experiment. Thus, C4.5/J4.8 classifier algorithm is considered as a potential classifier for future work. In this experiment, the generated classification rules from classification process can be used to predict the performance of employee to determine whether he/she has the possible for promotion or not. In conclusion, the ability to continuously change and obtain new understanding of the classification and prediction in HR research has become the major contribution to data mining in HRM.

REFERENCES

A TP Track Research Report (2005). *Talent Management: A State of the Art*: Tower Perrin HR Services.

Becerra-Fernandez, I., Zanakis, S. H., et al. (2002). Knowledge discovery techniques for predicting country investment risk. *Computers & Industrial Engineering, 43*(4), 787-800.

Chen, K. K., Chen, M. Y., et al. (2007). *Constructing a Web-based Employee Training Expert System with Data Mining Approach*. Paper presented at the 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).

Chien, C. F., & Chen, L. F. (2007). Using Rough Set Theory to Recruit and Retain High-Potential Talents for Semiconductor Manufacturing. *IEEE*

Transactions on Semiconductor Manufacturing, 20(4), 528-541.

Chien, C. F., & Chen, L. F. (2008). Data mining to improve personnel selection and enhance human capital: A case study in high-technology industry. *Expert Systems and Applications, 34*(1), 380-290.

Cubbingham, I. (2007). Talent Management : Making it real. *Development and Learning in Organizations, 21*(2), 4-6.

Delen, D., Walker, G., et al. (2005). Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligent in Medicine, 34*(2), 113-127.

ExecutiveBrief. (2008). 12 Competencies: Which Ones Should Your People Have. Retrieved 13/11/2008, from http://www.executivebrief.com.

Han, J., & Kamber, M. (2006). *Data Mining : Concepts and Techniques*. San Francisco: Morgan Kaufmann Publisher.

Huang, M. J., Tsou, Y. L., et al. (2006). Integrating fuzzy data mining and fuzzy artificial neural networks for discovering implicit knowledge. *Knowledge-Based Systems, 19*(6), 396-403.

Jantan, H., Hamdan, A. R., et al. (2009, 25-27 February 2009). *Knowledge Discovery Techniques for Talent Forecasting in Human Resource Application.* Paper presented at the World Academy of Science, Engineering and Technology, Penang, Malaysia.

Kumar, P. R., & Ravi, V. (2007). Bankruptcy prediction in banks and firms via statistical and intelligent techniques : A review. *European Journal of Operational Research, 180*(1), 1-28.

Ranjan, J. (2008). Data Mining Techniques for better decisions in Human Resource Management Systems. *International Journal of Business Information Systems, 3*(5), 464-481.

Tai, W. S., & Hsu, C. C. (2005). A Realistic Personnel Selection Tool Based on Fuzzy Data Mining Method. Retrieved 9/1/2008, from www.atlantis-press.com/php/download_papaer?id=46.

Tso, G. K. F., & Yau, K. K. W. (2007). Predicting electricity energy consumption : A comparison of regression analysis, decision tree and neural networks. *Energy, 32*, 1761 - 1768.

Tung, K. Y., Huang, I. C., et al. (2005). Mining the Generation Xer's job attitudes by artificial neural network and decision tree - empirical evidence in Taiwan. *Expert Systems and Applications, 29*(4), 783-794.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. San Francisco: Morgan Kaufmann Publishers.