Applied and Computational Mathematics
2019; 8(4): 75-81
http://www.sciencepublishinggroup.com/j/acm
doi: 10.11648/j.acm.20190804.12
ISSN: 2328-5605 (Print); ISSN: 2328-5613 (Online)

Prediction of Academic Talent Capacity Based on Gradient Boosting Decision Tree

Shunshun Shi, Mingzhou Chen*, Rui Feng, Hua Zhang, Shuai Zhang

School of Information, Zhejiang University of Finance and Economics, Hangzhou, China

Email address:

Shishunshun1993@126.com (Shunshun Shi), 569907429@qq.com (Mingzhou Chen), hzfengr@163.com (Rui Feng), zh_2792@163.com (Hua Zhang), zhangshuai@zufe.edu.cn (Shuai Zhang)

*Corresponding author

To cite this article:

Shunshun Shi, Mingzhou Chen, Rui Feng, Hua Zhang, Shuai Zhang. Prediction of Academic Talent Capacity Based on Gradient Boosting Decision Tree. *Applied and Computational Mathematics.* Vol. 8, No. 4, 2019, pp. 75-81. doi: 10.11648/j.acm.20190804.12

Received: August 5, 2019; **Accepted**: September 25, 2019; **Published**: September 27, 2019

Abstract: Talent introduction is an important force of academic development in universities. As the core of talent introduction, prediction of academic talent capacity is an essential and valuable research. However, it is hard to apply traditional statistical methods to extract knowledge from the mass and multi-dimensional talent information. Data mining approaches as up-to-date and efficient technologies are good at analyzing information, extracting patterns or rules from a big dataset and then making a prediction based on the relationship among extracted information. In this study, a series of data mining approaches are employed to evaluate the academic capacity of talent and to analyze the correlation between features. The Principal Component Analysis and Random Forest are used to feature extraction for improving the accuracy of prediction. A classical classification model, Gradient Boosting Decision Tree, is used as the primary analytic model to prediction. In order to validate the effectiveness of the model, other five classification models are used to conduct a comparative experiment based on prediction accuracy values and the F-measure metric. Further, to investigate the contribution of some important features, we make a marginal utility analysis of important features which have a high correlation with academic talent capacity. The experiment results reveals the important features for academic capacity and the positive factors for the academic production of talents.

Keywords: Data Mining, Classification Models, Prediction, Talent Introduction, Academic Talent Capacity

1. Introduction

Talent is an essential source of knowledge evolving and can foster the current and future performance of institutes [1], especially for universities. Talent introduction is an important part of university construction. Due to the increasingly strong economic growth, China universities invest more than ever in talent introduction. Several studies have showed that recent governmental policies in China have promoted the returning of talent from overseas [2-4]. Meanwhile, the development of talent in domestic universities has promoted the talent introduction to universities either. However, the increased talent number doesn't bring increased academic capacity definitely. Therefore, the academic capacity of talents should be evaluated quantitatively. It is found that the conventional evaluation approaches of academic talent capacity have no longer met the current requirements due to their static characteristics [5]. Thus, it is necessary to introduce a scientific evaluation approach of academic talent capacity for more complicated situations.

Data mining is one of the most useful approaches to analyze information, extract patterns or rules from a big dataset and predict the future trend of target label. Data mining is generally divided into five categories which are clustering, association, classification, prediction and outlier analysis. In this study, we use classification models to predict the academic talent capacity. We obtain a list of CV information of high-level talents that have been recruited by Zhejiang University of Finance & Economics, China in the last decade, and develop some classification-based models to predict their academic capacity based on their CV information. In this study, the academic capacity of a talent is evaluated by checking whether s/he obtains a Natural Science Foundation of China (NFC) in three years after s/he is recruited to the university. Gradient Boost Decision Tree [6], Decision Tree

[7], Random Forest [8], Artificial Neural Network [9], Naive Bayes [10] and Support Vector Machine [11] are used for prediction.

The remainder of this paper is organized as follow. Section 2 illustrates the related work about data mining models and their applications. Section 3 introduces the procedure of data pre-processing and modeling. Section 4 gives a detailed evaluation of comparative experiment based on prediction accuracy values and the F-measure metric. In the last section of this paper, a conclusion is provided including the limitations and future work of our study.

2. Related Work

Data mining measures are effective for extracting valuable information from a large dataset based on the data mining models, such as Support Vector Machines, Decision Tree, and Artificial Neural Network. Data mining has been applied in many fields varying from marketing, banking to health caring and education [12-15]. Recently, human resource field becomes an increasingly hot topic combined with data mining [16-18]. The applications of solving human resource problems based on data mining approaches have validated the feasibility and effectiveness of this combination [19-22].

The current human resource research based on data mining approaches mainly focuses on personnel selection which refers to putting right people to right position [23]. However, researches on predicting academic talent capacity based on talents' personal features are rare. In Chinese universities, obtaining a NFC is an important evaluation indicator of academic capacity for academic talents. Therefore, we choose this attribute as our evaluation indicator. In this study, we explore the relationship between talents' personal features and their academic capacity, where the academic capacity is abstracted into a label indicating whether a talent obtains a NFC in three years after s/he is recruited to the university. Besides, based on the extracted relationship, we have also discovered some valuable features which are important for the talent assessment. We use six classification models to analyze our dataset and among them, we can choose the model with highest prediction accuracy as our primary analytic model.

3. Method

In this paper, six classification models are used for predicting an academic talent whether s/he could obtain a NFC in three years after s/he is recruited to the university based on some selected features. We have obtained the raw dataset thanks to the support from Zhejiang University of Finance & Economics, China. However, the raw dataset contains some invalid data and irrelevant data, which would lead to unsatisfactory prediction results. Thus, we need to conduct a data preprocessing before using them to train the classification models. In order to improve the prediction accuracy as high as possible, we do a lot of data preprocessing work like removing invalid features, merging new features as well as removing non-correlated features based on Principal Component Analysis and Random Forest. Besides, all the experiments are implemented by R programming.

3.1. Data Collection

In this study, the main task is to evaluate the capability of the talents to obtain a NFC. The talent information comes mainly from their CV information and university database, covering 245 talents. We use the status whether a talent obtains a NFC as the dependent variable (i.e., prediction label) and the other features are used as independent variables.

3.2. Data Preprocessing

The raw data cannot be used in our prediction model directly. We need to preprocess the raw data for better predicting the academic capacity of talents. The features contained in each talent's CV are not unified, which means some features are contained in some talents' CV while are not in others'. Therefore, we need to exclude these features. Besides, some non-correlated features also should be removed and new indexes for synthesizing some low correlated features should be established.

The initial features are derived from the CV information and the university database. They include the status of obtaining a NFC and the number or the level of published papers, etc. Some features involve personal privacy or university confidentiality, so these features are handled through desensitization treatment or removed. On the other hand, the forms of some features in talent CV are not unified. If we use all features offered in all talents' CV, the evaluation result would be affected by the high dimensional data and the missing value of the whole features. To solve this problem, we retain the most common features for the subsequent analysis and remove the other invalid features.

There are some features belonging to a common category but classified into different hierarchies. We need to assign different values to these related features for differentiating their hierarchies. For example, the talents got their doctorates from different university schools. Some schools rank in international Times 100, some in international Times 150, some in China 985 project or China 211 project, while some are even not included in any of them. They are all schools from which the talents obtained their doctorates, and should be classified into different hierarchies. Therefore, we assign different values to these features and the detailed information is shown in table 1. Besides, there are some features that need to be integrated into a new comprehensive feature for better expressing its semantics. For example, talents have published SCI papers, SSCI papers, EI papers, etc. These papers all represent the talents' academic capabilities but SCI papers and SSCI papers are the highest levels of papers. Therefore, we choose the numbers of SCI papers and SSCI papers as two independent features, meanwhile, merge the numbers of all academic papers into a new feature to express the comprehensive academic paper quality. We use a new feature, CQP, to express it.

Table 1. *The description of 27 features in classification models.*

Features	Description
NSCI (FA)	The number of SCI papers (First author)
NSCI (SA)	The number of SCI papers (Second author)
NSCI (CA)	The number of SCI papers (Corresponding author)
NSCI	The number of SCI papers
DDGS	The degree of doctoral graduate school: (1) 3.5 represents schools ranking in USnews top 100 and international Times top 100. (2) 2.5 represents schools ranking in USnews top 150 and Times top 150. (3) 3 represents China C9 schools. (4) 2 represents schools sponsored by China Project 985. (5) 1 represents schools sponsored by China Project 211. (6) 0 represents the remaining schools.
T1	Period of abroad schooling during doctoral study
T2	Period of working in industry
T3	Period of schooling during post-doctoral education
T4	Period of other university teaching experience before entering the university
T5	Period of working in industries before entering the university
Age	The age of talents
Gender	The gender of talents: (1) 1 represents male. (2) 0 represents female.
MS	Marriage status: (1) 1 represents married. (2) 0 represents unmarried.
CQP	The comprehensive quality of papers.
NS	The doctoral degree program is natural science
Law	The doctoral degree program is law
Philosophy	The doctoral degree program is philosophy
Literature	The doctoral degree program is literature
Engineering	The doctoral degree program is engineering
Management	The doctoral degree program is management
Education	The doctoral degree program is education
Economics	The doctoral degree program is economics
History	The doctoral degree program is history
DUS	The degree of undergraduate school
NCA	The number of comprehensive awards
FD	Whether or not obtain a foreign degree
NFC	Whether or not obtain the NFC project in three years after employed by the university: (1) 1 represents yes. (2) 0 represents no.

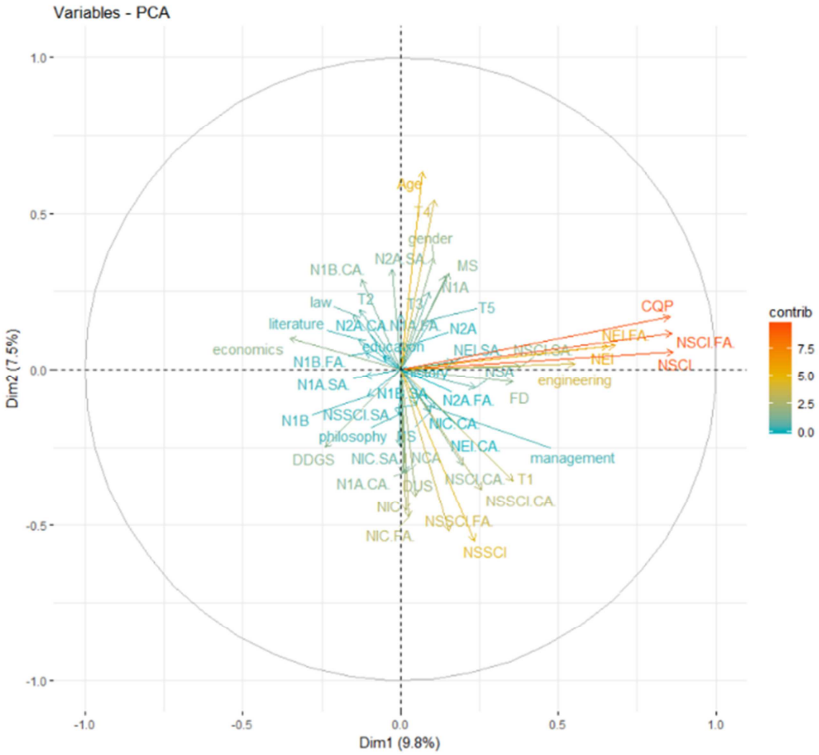

Figure 1. *The accuracy and F-measure metric of five classification models for BCCD data.*

In this section, we conduct feature extraction according to the correlation between features and information divergence of features. The process of feature extraction can improve the accuracy of prediction. The Principal Component Analysis and Random Forest are used to analyze the potential correlation. First, a covariance matrix which is calculated by

Principal Component Analysis can be used to analyze the correlation relationship among the features. From figure 1, we can find CQP, NSCI (FA), NSCI contribute most to the prediction label NFC. Second, Random Forest can be used for feature extraction either. According to Random Forest, we can identify some features with higher information divergence value. The higher the information divergence value is, the more powerful the classification capability of feature is. Therefore, the features with higher information divergence values are more inclined to be retained as our independent variables. The top five features with the highest information divergence values are shown in figure 2. Combining these two methods, we can conduct feature extraction and finally obtain 26 features for further investigation. Besides, some of the features are text type, such as "DDGS", "Gender", "MS" and "NFC", so we transformed them into numeric type. The detailed description of 26 independent variables and 1 prediction label is shown in table 1.

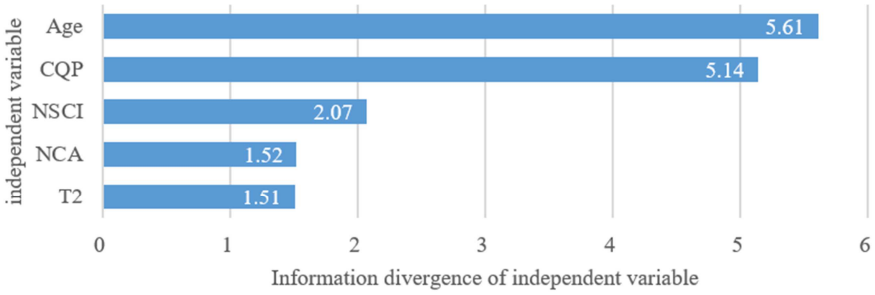

Figure 2. *A contribution of 26 features to NFC based on Principal Component Analysis.*

3.3. Modeling

In this section, we use six classification models to predict whether or not a talent could obtain NFC in three years, then choose the model with the highest prediction accuracy as our primary model. The features in table 1 except NFSC are the independent variables. Based on these features, we investigate the relationship between them and the prediction label NFC. A 10-fold cross validation is used for deriving better prediction accuracy.

In this study, we use the Gradient Boosting Decision Tree (GBDT) model as the basic model among six alternative models. We introduce the advantages of GBDT concisely. One the one hand, the GBDT can handle multiple types of data, which can be efficiently applied in this project. On the other hand, there are various types of loss functions ensuring the robust of GBDT. A flow chat of GBDT is provided in figure 3. The GBDT algorithm emphasizes on a boosting technology. In each iteration, the loss function is calculated according to the residual of predicted label values and practical label values. Then, it will produce a new Classification and Regression Tree to fit the residual and the GBDT is updated. In this project, a few CVs have unified format, which causes information loss. For example, the feature T2 is not presented in some CVs. In addition, there are a small number of talents that were recruited in special period, which would be considered as abnormal points. The GBDT consisted of a large amount regression trees which extract of different features and have different structures and gradient boosting technique can overcome above impediments. According to the experiment results, we derive that the prediction accuracy of GBDT is the highest compared with the other five models. Therefore, we choose GBDT as our primary model.

Figure 3. *The flow chart of GBDT model.*

4. Experimental Results

This section consists of two parts. One is an evaluation of comparative experiment based on six classification models, and the other is a marginal utility analysis of important features, which further investigates the contribution of some important features.

4.1. Experiment Design

For better evaluating the performance of six classification models, we use a 10-fold cross validation to process our experiment. The dataset is divided into 10 groups with similar size, and every group is used for testing the classification precision of the remaining 9 groups. Finally, the prediction accuracy and F-measure metric values of six classification models are derived as shown in figure 4 and table 2. The F-measure metric is the harmonic average of precision and recall, so it considers both the performance of precision and recall. The higher the F-measure is, the more effective the model is. We apply six classification models to analyze the

dataset, which are Random Forest (RF), Decision Tree (DT), Artificial Neural Network (ANN), Naive Bayesian (NB), and Support Vector Machine (SVM). The GBDT model ranks first both in accuracy and F-measure metric. Therefore, GBDT model is selected as our primary analytic model.

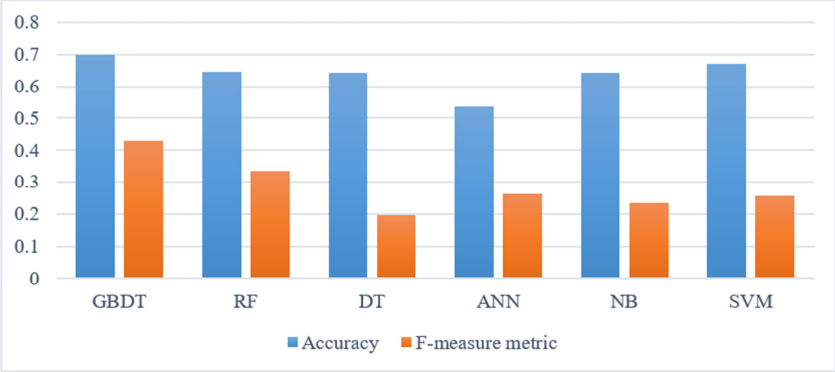

Figure 4. *The accuracy and F-measure metrics of six classification models.*

Table 2. *The detailed information about F-measure metric of six classification models.*

Classification model	GBDT	RF	DT	ANN	NB	SVM
Accuracy	0.700	0.647	0.641	0.535	0.641	0.671
F-measure metric	0.430	0.335	0.196	0.261	0.235	0.255

4.2. Experiment Evaluation

In the section of preprocessing, we notice that a few features have a bigger impact on NFC than the others, so in this subsection, we calculate the correlation coefficient of 26 independent features with NFC based on the GBDT model. Then we select two features with the highest correlation coefficient and conduct a marginal utility analysis between them and NFC to further investigate the relationship among them.

From the correlation coefficient values of the 26 features, we can find that Age, CQP and NSCI rank first, second and third respectively among the total 26 features. The higher the correlation coefficient value of feature is, the closer the relationship between the feature and NFC is. Therefore, these three features are the most relevant to NFC. In other words, these three features are the most important for measuring academic talent capacity. Therefore, we conduct a marginal utility analysis for exploring how they contribute to NFC. However, CQP is a comprehensive feature which considers the published papers with different levels. Thus, it cannot be compared with other features due to its high-dimensional characteristic and we decide to choose Age and NSCI as the features for marginal utility analysis. From figure 4, we find NSCI has positive relationship with NFC and the marginal utility of NSCI is getting stronger with the increase of NSCI value. It indicates the more SCI papers a talent has published, the higher his/her academic capacity is. However, Age has negative relationship with NFC and the marginal utility of Age is negative when the age is higher than 35. It demonstrates that if a talent is employed at the age of 35 or older, his/her academic capacity in obtaining NFC tends to be decreased. By combining these two features, we can infer that younger talents who have published more SCI papers are the academic flesh troops and they have higher capability to obtain NFC. The talent introduction of university should be paid more attention to these people.

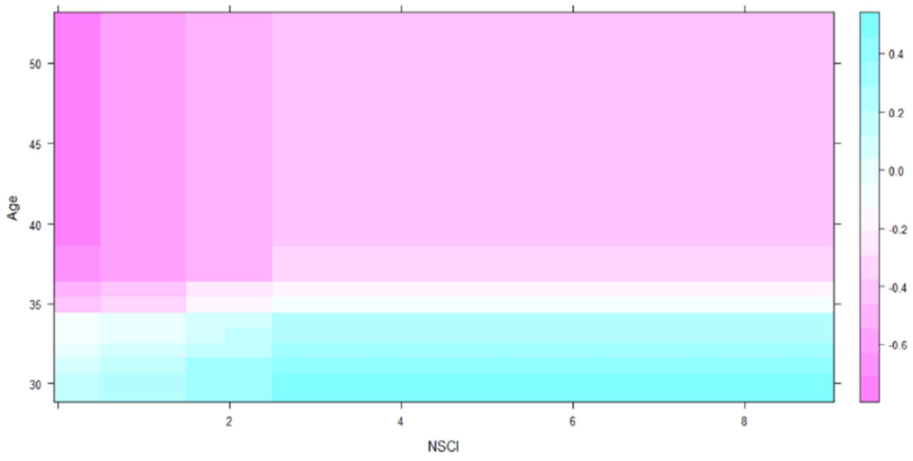

Figure 5. *Two-dimension marginal utility of Age and NSCI corresponding to NFC.*

5. Conclusion

In this study, some classification models are explored in the application of academic talent introduction in universities. Before applying the dataset to train the classification models, we do a lot of data preprocessing work in order to achieve the better prediction result. Then we use the Gradient Boosting Decision Tree (GBDT) model to analyze our dataset. In order to validate the effectiveness of the GBDT model, we use another five classification models to conduct the comparative experiments. The experiment result has proved that GBDT model is the best option for exploring the relationship between the 26 selected features and the NFC. According to the relationship we have discovered, some important independent features can be identified. Age of talents and number of SCI papers talents have published are the two features that contribute to their academic capacity mostly. Younger talents with more published SCI papers tend to be more prolific in academic job and have higher capability to obtain academic achievements such as NFC projects. Thus, this study provides a guide for the talent introduction of universities. Younger talents with more published SCI papers are preferred in academic talent introduction. Besides, other features also positively affect the academic production of talents, like the number of SSCI papers, the number of SCI paper (FA), and the number of comprehensive awards, etc.

However, this study still has some limitations. One is related to the data mining model. Compared with Random Forest, Decision Tree, Artificial Neural Network, Naive Bayesian, and Support Vector Machine, it has been proved that GBDT is the best model to analyze our dataset. But, the prediction accuracy value of GBDT model is not high enough and still needs to be promoted. Therefore, further investigations on extending algorithms should be conducted in future. Another drawback is related to the size of dataset. In this study, the size of dataset is not big enough. Some underlying relationship among the talents' CV information and their academic capacity has not been discovered yet. We believe that more valuable relationships and information could be revealed if the dataset could be expanded.

References

[1] Hanif, M. I. & Yunfei, S. (2013), *The role of talent management and HR generic strategies for talent retention*, African Journal of Business Management, 7, 2827-2835.

[2] Kellogg, R. P. (2012), *China's brain gain: Attitudes and future plans of overseas Chinese students in the US*, Journal of Chinese Overseas, 8, 83-104.

[3] Tharenou, P. & Seet, P. S. (2014), *China's reverse brain drain: regaining and retaining talent*, International Studies of Management and Organization, 44, 55-74.

[4] Ma, Y. P. & Pan, S. Y. (2015), *Chinese returnees from overseas study: An understanding of brain gain and brain circulation in the age of globalization*, Frontiers of Education in China, 10, 306-329.

[5] Lievens, K. van Dam, & Anderson, N. (2002), *Recent trends and challenges in personnel selection*, Personnel Review, 31, 580-601.

[6] Friedman, J. H. (2001), *Greedy function approximation: A gradient boosting machine*, Annals of Statistics, 29, 1189-1232.

[7] Quinlan, J. R. (1987), *Simplifying decision trees*, International Journal of Man-machine Studies, 27, 221-234.

[8] Breiman, L. (2001), *Random forests*, Machine learning, 45, 5-32.

[9] Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996), *Artificial neural networks: A tutorial*, Computer, 29, 31-44.

[10] Chen, J., Huang, H., Tian, S., & Qu, Y. (2009), *Feature selection for text classification with Naive Bayes*, Expert Systems with Applications, 36, 5432-5435.

[11] Suykens J. A. & Vandewalle, J. (1999), *Least squares support vector machine classifiers*, Neural Processing Letters, 9, 293-300.

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

[13] Hormozi, A. M. & Giles, S. (2004), *Data mining: A competitive weapon for banking and retail industries*, Information Systems Management, 21, 62-71.

[14] Koh, H. C. & Tan, G. (2011), *Data mining applications in healthcare*, Journal of Healthcare Information Management, 19, 65-72.

[15] Romero, C. & Ventura, S. (2013), *Data mining in education*, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3, 12-27.

[16] 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 with Applications, 34, 280-290.

[17] Ranjan, J., Goyal, D. P. & Ahson, S. I. (2008), *Data mining techniques for better decisions in human resource management systems*, International Journal of Business Information Systems, 3, 464-481.

[18] Gupta, S., Mokashi, U. M., & Suma, V. (2017). *Entropy-based discretisation for performance prediction of employee: strategy for improving software quality*, International Journal of Productivity and Quality Management, 21, 411-428.

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

[20] Han, Y. (2016). *Improved BIRCH Clustering Algorithm and Human Resource Management Efficiency: An Organizational Learning Perspective*. International Journal of Security and Its Applications, 10 (8), 385-394.

[21] Fadhil, R., Djatna, T., & Maarif, M. S. (2017). *Analysis and Design of a Human Resources Performance Measurement System for the Nutmeg Oil Agro-industry in Aceh*. Journal of Regional and City Planning, 28 (2), 99-110.

[22] 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, 528-541.

[23] Saron, M. & Othman, Z. A. (2012), *Academic talent model based on human resource data mart*, International Journal of Research in Computer Science, 2, 29-35.