Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique

Pornthep Khongchai
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
p.khongchai.22@gmail.com

Pokpong Songmuang
Department of Computer Science
Faculty of Science and Technology
Thammasat University
Pathumthani, Thailand
pokpong@cs.tu.ac.th

Abstract—**This paper presents a salary prediction system using a profile of graduated students as a model. A data mining technique is applied to generate a model to predict a salary for individual students who have similar attributes to the training data. In this work, we also made an experiment to compare five data mining techniques including Decision trees, Naive Bayes, K-Nearest neighbor, Support vector machines, and Neural networks to find the suitable technique to the salary prediction. In the experiment, 13,541 records of graduated student data were used with 10-fold cross validation method. Results showed that K-Nearest neighbor provided the best efficiency to be used as a model for salary prediction. For usage evaluation, a questionnaire survey was conducted with 50 user samplings and a result showed that the system was effective in boosting students' motivation for studying and also gave them a positive future viewpoint. The result also informed that they found they satisfied with the implemented system since the system was easy to use, and the prediction results were simple to understand without requiring any background knowledge.**

Keywords-Motivation; Salary prediction system; Educational data mining; Classification technique;Decision trees; Naive Bayes; K-Nearest neighbor; Support vector machines; Neural networks

I. INTRODUCTION

Nowadays, most of university students take courses without determination or career goal since they mostly choose a field of study following their friends or trend. With a goal in life, they lack motivation to study and often sway to surrounding temptation such as social activity or entertainment. This causes them to accumulate their boredom in studying in which leads to a failure in studying lessons and a bad performance in the examinations [1]. This problem moreover results in students' dropping out from university by either resignation or getting retirement.

To increase their motivation, a good example of successful graduated students should be demonstrated. The graduated students are an example of a person who already took the same path as the current students do. They who have a good career and constant income can help as a role model to motivate current students in studying, gaining their career plan or even setting their goal.

One of the factors being concerned in planning the future is an income. Using income rate to motivate studying becomes common in advising a career plan. However, a rate of an income can vary based on several factors such as a famous of university name, students' GPA, and activities students do while studying. Therefore, to use an income in motivating students should only apply the exampling persons from the local university. To predict a salary, a history in studying of the graduated students is used as a model to reflect factors for a different rate of salary. With the model, a salary prediction system [2][3] can be implemented as a tool to show an example of previous people with a successful career along with activities that they conducted while being a student.

Many previous studies proposed salary prediction models and systems [2][3][4] using regression technique. Although these models performed well, some problems were noticeable. The problems include 1) the models predicted salaries for a group, not for individual students, and 2) the results from the prediction models required an extensive background in statistics to fully comprehend.

Therefore, this research proposed a model to predict students' future salaries based on graduated student history. By inputting students' activities and grades in individual, salary of each student can be predicted, and the result of the prediction is a salary rate. In this work, data mining techniques are exploited for the salary prediction model. Several data mining techniques have been tested to compare which is the best in the task.

This paper describes the literature review in Section 2. The proposed system is explained in detail in Section 3. Section 4 presents the experimental results, including accuracy and efficiency. Finally, Section 5 provides a conclusion with a summary of the findings

II. LITERATURE REVIEW

This paper aims to determine efficient data mining techniques [5] for predicting salary to motivate students during their studies. Several data mining techniques previously employed in the educational field were reviewed.

Young-joo Lee et al. [2] applied regression methods to predict both salary and job satisfaction based on the historical

978-1-5090-5130-4/16/$31.00 ©2016 IEEE

data of graduated students. Jerrim [3] proposed an ordinary least-squares (OLS) regression model to construct prediction models of future salaries for American college students based on family background data, and social profiles. Karla et al. [4] used hierarchical linear regression to construct a prediction model with student and program characteristics as control variables, and salary as the predictor variable. However, some problems included 1) the prediction model only predicted salaries for a group, not for individual students, and 2) the results from the prediction model required extensive statistical knowledge to fully comprehend.

Maomi Ueno [6][7] proposed a pedagogical agent, based on a decision tree model constructed from learning historical data to predict a student's future final status into four groups as 1) Failed, 2) Abandon, 3) Successful, and 4) Excellent. Farhana Sarker et al. [8] introduced a decision tree model to predict the marks of students, based on internal institutional and external open data sources used in practical settings to predict academic performance.

C. Márquez-Vera et al. [9] predicted which factors dominated the dropout rate by analyzing data concerning 670 middle-school students from Zacatecas, México using data mining techniques. The data were collected from university and college databases, and they applied various techniques to predict useful knowledge.

Previous studies indicated that data mining techniques were the most popular applications for prediction models in educational research. Moreover, these applications show an acceptable result and positive effect in using in an education field.

However, data mining techniques have never been applied for student salary prediction. Therefore, the performances of data mining techniques for salary prediction will be compared to determine the most suitable for creating a salary prediction system.

III. Salary Prediction System

The aim of a salary prediction system is to determine a salary of students based on activities they do while they are in university comparing to graduated student. We expect that the predicted salary will urge them to pay more attention on studying since their cloudy future becomes clearer in direction and outcome. Data mining technique is a core of the prediction system. Data of graduated students with their salary are used to train a salary-activity model, and the model will be used to predict a salary for current students.

A. System Design

The system requires two input data as follows:

1) Student profiles

2) Profiles of former students who graduated and their salary

In this work, three profiles of the former students who are mostly similar to the student and gaining highest salary are exemplified as exemplars. In this work, feature selection is applied to scale down a number of features to only the features significantly effect the prediction. The selected features to be compared are including:

- Gender (Male, Female)
- Faculty (Engineering, Business, Arts, etc.)
- Program (Engineering Civil, Computer Science, Marketing, etc.)
- Job Training (Yes, No): does a student apply for job training?
- Certificate (Yes, No): does a student receive a certification in their study related field?
- GPA (>2.79, <2.8)
- Salary (four levels as classes: less than 13,500, 13,501 - 15,300, 15,301 - 18,000, and more than 18,000 baht.): prediction result for current students and label from the graduated students

B. Model Training

Since there are many data mining techniques that can be applied for salary prediction, we select five models to compare in this study as follows:

K-Nearest neighbors (K-NN) is an algorithm of a classifier measured by a distance vector, and has been used to evaluate many algorithms. This experiment used IBk algorithm which determines weights of distance based on cross-validation [10].

Naive Bayes (NB) [11][12] is a type of classifier based on probability. This method applied Bayes' theorem with strong independence assumptions between the features and weighted naive Bayes to fit for training and classification.

Decision trees (J48) [13] represent the hierarchical nature of a structure by node graphs using WEKA. This algorithm was applied to compare all techniques used in this experiment.

Multilayer perceptron (MLP) [14][15] algorithm is an extensively used and popular neural network. This research used MLP to evaluate the model for comparison.

Support vector machines (SVM) [16][17] is a technique to separate features by kernel function, and it was also applied here.

The efficiencies of the above data mining techniques will be compared for predicting salary, and the best one will be installed in the system.

C. Usage of Salary Prediction System

The goal of this system is to show a predicted salary to increase students' motivation. The system motivated the student using 1) the salary prediction from student profiles, and 2) the examples of three graduated students with similar profiles and top-ranked salaries. Accordingly, users need to be able to understand the predicted results without the necessity of a statistical background.

For an input of the system, students are asked to fill in their information as profile. A user interface for filling information is shown in Figure 1.

Figure 1 illustrates the interface of the salary prediction system. The interface contained the seven attributes asking to be filled in. In this example, they are gender, faculty, program, type of work, job training, certificate, and GPA. After data are input and submitted, the system compared the inputs of attributes with

the rules and displayed the predicted salary of the three graduated persons whose attributes are mostly similar to the input and having highest salary rate as shown in Figure 2.

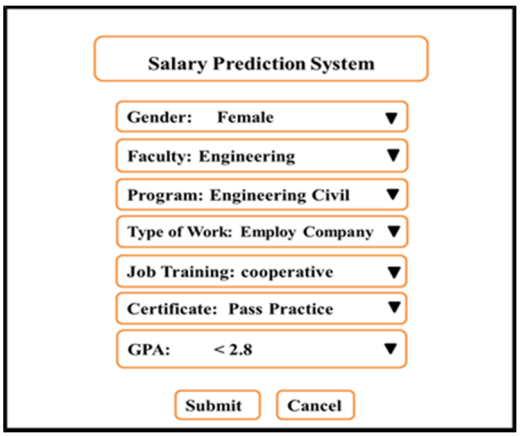
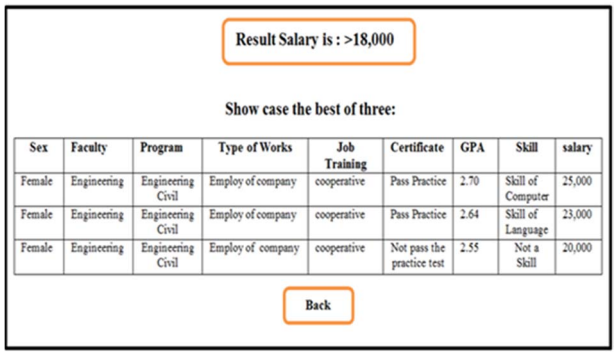


Figure 1: Interface of the Salary Prediction System



Sex	Faculty	Program	Type of Works	Job Training	Certificate	GPA	Skill	salary
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.70	Skill of Computer	25,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Pass Practice	2.64	Skill of Language	23,000
Female	Engineering	Engineering Civil	Employ of company	cooperative	Not pass the practice test	2.55	Not a Skill	20,000

Figure 2: Salary Prediction Result

In Figure 2, the predicted salary is more than 18,000 baht. Moreover, the salary prediction model compared the inputs of attributes with the profile in the database, and selected three graduated students with similar attributes with the highest salaries. The system presented profiles of these three graduated students as follows: gender, faculty, program, type of work, job training, certificate, GPA, skill, and salary.

The proposed salary prediction system show the potential to overcome the problems of previous systems as 1) the salary prediction system can predict salaries of individual students, and 2) the users do not require statistical knowledge to understand the results from the salary prediction system.

IV. Experiment and Results

To select the best data mining technique for using in the salary prediction system, we set up an experiment to compare six techniques mentioned in Section 3.A. Data in testing are a data of students who graduated in 2006 to 2015 from Rajamangala University of Technology Thanyaburi, Thailand. The historical data included gender, student faculty, student program, type of work, experience of job training, received certification, total Grade Point Average (GPA), salary, address, telephone number, and e-mail. In total, there are 13,541 profile records in use for a training data.

A. Data Preparation

Data of graduated students have accumulated over 10 years, and the base salaries have increased relatively each year based on an economic value. Therefore, former salaries cannot be directly used to create a salary prediction model for recent students. Thus, linear equating techniques were applied to adjust previous salaries to current salaries, based on the hypothesis that the distribution of salaries was similar for every year. Firstly, step-wise data preparation was conducted, and salary prediction models were then created and used in the salary prediction system.

Moreover, the data of graduated students contained several missing values and excessive numbers of attributes for creating the salary prediction model. In addition, the data format required reformation since they are apparently unacceptable for the data mining tool. Data was therefore prepared following three steps:

1) Data Selection: Relevant attributes were selected for salary prediction and forward selection [18] analysis was used to select 7 out of 108 total attributes. These 7 attributes were as follows: gender, faculty, program, type of work, job training, certificate, and GPA.

2) Data Cleaning: Outlier data (e.g. noticeably high salary) and missing values (e.g. no salary records) were manually removed. After data cleaning, the remaining data comprised 13,541 rows.

3) Data Transformation: This step prepared the data format as usable for the data mining tool. The salary data was changed from continuous to an interval for-mat. User specific discretization [19] was applied, and the salaries were divided in-to four levels as classes: less than 13,500, 13,501 - 15,300, 15,301 - 18,000, and more than 18,000 baht.

B. Comparison of Data Mining Techniques

Data mining techniques were compared for predicting salary using the data mining tool WEKA (Waikato Environment for Knowledge Analysis) version 3.6 [20]. A 10-fold cross-validation technique was employed to evaluate the efficiency of the salary prediction model created using data described in the previous section. Table 1 presents the results of the salary prediction models in terms of Recall, Precision, F-measure, and Overall Accuracy [21][22].

The prediction of each data mining technique is shown in Table 1. For Recall, Precision, and F-measure the highest overall accuracy prediction was 84.69% for KNN, and the lowest was 38.08% for MLP. Decision trees (J48) had a percentage of 73.96%, Support vector Machines (SVM) (43.71%), Naive Bayes (NB) (43.63%), and Multilayer perceptron (MLP) (38.8%). KNN had the highest overall prediction accuracy results of the run models. Recall had the best results on classes of salary with more than 18,000 baht scoring 87.10%. KNN had the highest prediction results after run model. Precision had the best result on salary with more than 18,000 baht scoring 86.30%, and KNN had the highest F-measure results. F-measure showed high results for the class of salary with more than 18,000 baht scoring 86.70%.

For a small number of features, KNN showed higher efficiency than the other techniques and was also easy to

implement. On the other hand, with a larger number of features more complex techniques are required. K-NN showed the best overall accuracy; hence, a model from K-NN is selected to apply in the salary prediction system.

TABLE I. SUMMARY OF SALARY PREDICTION MODELS

Class (baht)	Recall (%)				
	KNN	NB	J48	MLP	SVM
Less than 13,500	**84.90**	50.90	75.80	45.30	37.00
13,501 - 15,300	**83.40**	49.70	71.80	41.00	58.00
15,301 - 18,000	**83.70**	21.70	71.80	0.50	23.60
More than 18,000	**87.10**	53.40	76.90	68.70	56.70
Class (baht)	**Precision (%)**				
	KNN	NB	J48	MLP	SVM
Less than 13,500	**84.90**	43.90	74.60	33.30	48.70
13,501 - 15,300	**84.10**	38.70	73.40	33.10	37.10
15,301 - 18,000	**83.60**	45.10	71.50	38.30	43.00
More than 18,000	**86.30**	49.30	76.50	47.70	51.00
Class (baht)	**F-measure (%)**				
	KNN	NB	J48	MLP	SVM
Less than 13,500	**84.90**	47.10	75.20	38.40	42.10
13,501 - 15,300	**83.70**	43.50	72.60	36.70	45.30
15,301 - 18,000	**83.60**	29.30	71.70	1.00	30.50
More than 18,000	**86.70**	51.30	76.70	56.30	53.70
Accuracy (%)	**84.69**	**43.63**	**73.96**	**38.08**	**43.71**

C. Evaluation of Student Motivation

A questionnaire was designed in three parts to evaluate student motivation. The samples were 50 students from Rajamangala University of Technology Thanyaburi, Thailand. The same source used for the salary prediction system.

a) Part 1: General student information

Basic student information was collected including gender, age, faculty, and program. In summary, there were 26 males and 24 females, with ages ranging from 19 to 22. The students came from 7 different faculties, and numbered 7, 15, 5, and 23 from years 1 to 4, respectively.

b) Part 2: Questions to examine student motivation before using the system.

The questions and results are shown in Table 2.

TABLE II. QUESTIONS BEFORE USING THE SALARY PREDICTION SYSTEM AND RESULTS

Question No.	Question	Scores / Answers	Mean(SD)/Full scores
Q1	Does your current faculty/program match your objectives?	1.Yes, 2.No	1.26 (0.44) / 2
Q2	Which level of satisfaction do you have for your faculty/ program?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	3.8 (0.57) / 5
Q3	Does a high salary affect your study motivation?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.16 (0.47) / 5
Q4	Which salary rate do you expect after graduation?	1. Least than 13,500 2. 13,501 - 15,300 3. 15,301 - 18,000 4. More than 18,000	3.7 (0.34) / 4

c) Part 3: Questions to examine student motivation after using the system

The questions and results are shown in Table 3.

TABLE III. QUESTIONS AFTER USING THE SALARY PREDICTION SYSTEM AND RESULTS

Question No.	Question	Scores / Answers	Mean(SD)/Full scores
Q5	How much does the predicted salary from the system match your expected salary?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.00 (0.45) / 5
Q6	How much satisfaction do you have with your current faculty/program?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.16 (0.55) / 5
Q7	After knowing the predicted salary, how much satisfaction do you have with your enrolled faculty/program?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.12 (0.59) / 5
Q8	How much does the system motivate you to study for your expected salary?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.18 (0.40) / 5

The majority of answers from Q1 showed that most students were on track to achieve their objectives in their current faculty and program. The mean score results in Q2 showed that students were greatly satisfied with their current faculty and program. Moreover, combining Q2 with Q6, the predicted salary has an effect to increase students' satisfaction with their current faculty and program. Q7 results indicated that the salary predicted by the system made students to realize the implicit benefit of being in the current faculty and program, and improved their desire to

accomplish their own future plans. Q8 results inferred that knowledge of predicted salary gave the students a much clearer goal and motivated them to study to achieve their expected salary.

The salary prediction system was evaluated using a questionnaire survey. Questions 9 and 10 are shown in Table 4.

TABLE IV. AFTER USING THE SALARY PREDICTION SYSTEM

Question No.	Question	Scores / Answers	Mean(SD)/Full scores
Q9	Is the system easy to use?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	4.02 (0.40) / 5
Q10	Is the system's result reliable in your opinion?	1. Least, 2. Little, 3.Moderate, 4.Much, 5.Most	3.89 (0.38) / 5

Questions 9 and 10 evaluated the satisfaction from using the system. Results showed that the users were very satisfied with the designed user interface as it was easy to use. For the reliability of the output, the scores were distinctively above average, but not highly trustworthy. This result was understandable since future predictions are very hard to prove.

Moreover, Question 11 also asked for other suggestions related to the system. The suggestions were summarized into three topics as follows:

- Some students suggested that this system should be used before the entrance examination. They could then choose their academic path and goal rationally before making a decision on faculty and program.
- Some students noted that the prediction increased activity and motivation for studying as they could better see their future. Moreover, some mentioned that results from graduated students were useful as an exemplar to them.

Positive comments

1. This system has the concept of guiding students to plan their studies successfully to achieve good salaries.
2. The system assists students to logically choose courses to attain high salaries.
3. The system can be utilized before entering study as a guide for students to choose faculties and fields of study that will generate good incomes in the labor market.
4. The system can be utilized continuously to deliver information during study time.
5. In each faculty/field of study, the system can be utilized to consider additional characteristics such as English skills for higher income opportunities.
6. The system presents good information to individual graduates and is easy to understand.
7. The system can be used as information for parents regarding faculties/fields of study, and enhance comprehension and objectives of study regarding careers and incomes between parents and students.

Negative comments

1. Names/titles of careers or positions should be clearly identified as these are still unavailable in the system.
2. Failures/errors for salary predictions still occur which makes the system unreliable.
3. The current system cannot be adjusted for changes in salaries, resulting in prediction errors.

V. SUMMARY

This paper presents a salary prediction system using data mining technique. The system is designed to support individual based salary prediction by comparing profiles of current student and graduated students. We also compared data mining techniques that performed best in the task. An experiment was conducted using 13,541 records of graduated student data by 10-fold cross validation. Results indicated that K-NN gave the best accuracy at 84.69% while Multilayer perceptron returned the lowest accuracy at 38.08%. From the usage evaluation from 50 sampling users, they indicated that the system is helpful to increase their motivation in studying and made them realize their plan to achieve their goal. Moreover, they mentioned the easiness in system usage, and the prediction result was simple and comprehensible.

ACKNOWLEDGEMENT

The authors express their thanks to Office of Academic Resource and Information Technology, Rajamangala University of Technology Thanyaburi, Thailand for providing the dataset of graduated students.

REFERENCES

[1] Lumsden and Linda S., "Student Motivation To Learn". ERIC Digest, Number 92, 1994.

[2] Young-joo Lee and Meghna Sabharwal, "Education–Job Match, Salary, and Job Sat-isfaction Across the Public, Non-Profit, and For-Profit Sectors: Survey of recent college graduates", Public Management Review, 18:1,p 40-64, 2014.

[3] John Jerrim, "Do college students make better predictions of their future income than young adults in the labor force?", Education Economics, 23:2, p 162-179, 2013.

[4] Karla R. H. and W. A. Hamlen, "Faculty salary as a predictor of student outgoing salaries from MBA programs", Journal of Education for Business, 91:1, p 38-44, 2015.

[5] Romero C. and Ventura S., "Educational Data mining: A Review of the State of the Art", IEEE Transactions on Systems., Man and Cybernetics. 40(6), p 601-618, 2010.

[6] Maomi U. (2005) Animated Pedagogical Agent based on Decision Tree for e-Learning", Proc.IEEE conference (Computer Science), ICALT

[7] Maomi U. (2004) Animated agent to maintain learner's attention in e-learning, Proc. E-Learn

[8] Farhana S. et al.,"Students' Performance Prediction by Using Institutional Internal and External Open Data Sources", CSEDU, 2013. page 639-646. SciTePress, 2013.

[9] Carlos Marquez-V. et al.,"Predicting School Failure and Dropout by Using Data Mining Techniques", IEEE journal of Latin-American learning technologies, vol. 8, no. 1, p.7-14, 2013.

[10] T. Cover and P. Hart., "Nearest neighbor pattern classification", IEEE Transactions on Information Theory, 13(1):21–27, January 1967.

[11] Titterington D.M., Murray G.D., Murray L.S., Spiegelhalter D.J., Skene A.M., Habbema J.D.F., and Gelpke G.J., "Comparison of discrimination techniques applied to a complex data set of head injured patients", Journal of the Royal Statistical Society, Series A, 144, 145–175, 1985.

[12] Mani S., Pazzani M.J., and West J., "Knowledge discovery from a breast cancer database", Lecture Notes in Artificial Intelligence, 1211, 130–133, 1997.

[13] Quinlan, J. R., "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.

[14] S. German, E. Bienenstock, and R. Doursat, "Neural networks and the bias/variance dilemma", Neural Computation, 4(1):1-58, 1992.

[15] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all", Artificial Intelligence, 137(1-2):239–263, 2002.

[16] V.Vapnik, "The nature of statistical learning theory", Springer, New York, 1995.

[17] V.Vapnik. "Statistical Learning Theory", Wiley, 1998.

[18] Mark A. Hall and Geoffrey H., "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.15, NO. 3, 2003.

[19] Petr B. and Ivan B., "Discretization and grouping: preprocessing steps for Data Mining", Principles of Data Mining and Knowledge Discovery, Second European Symposium, PKDD '98, Nantes, France, 1998.

[20] Machine Learning Group at the University of Waikato, http://www.cs.waikato.ac.nz/ml/weka/downloading.html

[21] Osiris V., "Feature Selection and Classification Methods for Decision Making: A Comparative Analysis", College of Engineering and Computing Nova Southeastern University, 2015.

[22] George F., "An Extensive Empirical Study of Feature Selection Metrics for Text Classification", Journal of Machine Learning Research 3, p 1289-1305, 2009.