A Support Vector Regression-based Prediction of Students' School Performance

Jui-Hsi Fu[1], Jui-Hung Chang[2]
[1]Department of Computer Science and Information Engineering
[2]Department of Engineering Science
[1]National Chung Cheng University
[2]National Chung Kung University
[1] Chiayi County, Taiwan
[2]Tainan, Taiwan
e-mail: [1]fjh95p@cs.ccu.edu.tw
[2]changrh @mail.ncku.edu.tw

Yueh-Min Huang[3] and Han-Chieh Chao[4]
[3]Department of Engineering Science
[4]College of Electrical Engineering and Computer Science
[3]National Chung Kung University
[4]National Ilan University
[3]Tainan, Taiwan
[4]Ilan, Taiwan
e-mail: [3]huang@mail.ncku.edu.tw
[4]hcc@mail.niu.edu.tw

Abstract—**The relationship between a person's personality and performance has long been studied by psychologists. Research suggests that a person's performance and behavior are related to personality characteristics and background data to a certain degree. In this paper, the Big Five personality model is adopted for measuring profiles of students, whose undergraduate performance and behavior are then analyzed. A machine learning approach, support vector regression (SVR), is employed to find correlations from the given sample data. The performance and behavior of a person are predicted from the obtained regression values. Personality, biological, performance, and behavior data of 120 undergraduates in Taiwan were collected through questionnaires. Ninety valid data samples are used for training in SVR and the others are used for evaluating the regression predictions. Most of the predicted performance yielded near 80% accuracy. It is shown that there are correlations between a person's performance and personality characteristics. SVR is shown to be a suitable method for exploring personality correlations.**

Keywords-Big Five personality model; SVR; personality; performance

I. Introduction

School administrators consider academic achievement when recruiting students. Admission examinations such as the SAT in the US and the *General Certificate of Education* in the UK are commonly used for determining students' knowledge and learning ability, and thus are used by schools to predict students' academic achievement. In addition to course grades, aspects such as special abilities and behavior are considered. Students with outstanding skills in areas such as music and sport can increase the school's reputation. Students may have some unwanted behaviors such as tardiness and aggressiveness, or good behaviors such as social participation in student societies or clubs. Therefore, it is important to identify or predict such behavior during the student admission process.

Two *College Board*'s research reports [2, 3], which sampled vast amounts of college student data, found that SAT scores and high school records were good predictors of academic achievement in college. They also collected data and performed statistical analysis for other factors, such as gender and course types, and predicted performance in areas such as artistic skill, athletic ability, and leadership.

Many methods have been proposed for predicting students' learning performance. For example, neural networks were used to predict the possibility of a student going to college [4], the work performance of graduates [5], and the exam results of students who had taken a ten-months e-learning course [6]. The prediction results obtained using neural networks and other methods have been compared [7]. In addition to neural networks, methods such as recommendation systems for educational data mining and student performance prediction [8], a decision tree approach for predicting students' performance in an engineering dynamics course [9], and a data mining technique for predicting students' final exam results in an e-learning course [10], have been proposed.

In social science studies, correlations between factors and performances are explored through statistical methods, commonly linear regression [17]. Most studies establish one-to-one correlations [18]. However, to determine the correlations or relationships between multiple factors and performances, which are nonlinear patterns, more sophisticated regression methods must be used. This process is highly manual and tedious, requiring computation, analysis, and interpretation. The correlation combinations can be too complicated for human experts to apply in practice. Figure 1 shows examples of one-to-one correlations and multiple correlations, where r represents the Pearson's correlation coefficient. Most studies have presented one-to-one correlations, but performance is likely associated with more than one factor. It is not feasible for human experts to evaluate and predict performance using these data.

978-0-7695-4655-1/12 $26.00 © 2012 IEEE
DOI 10.1109/IS3C.2012.31

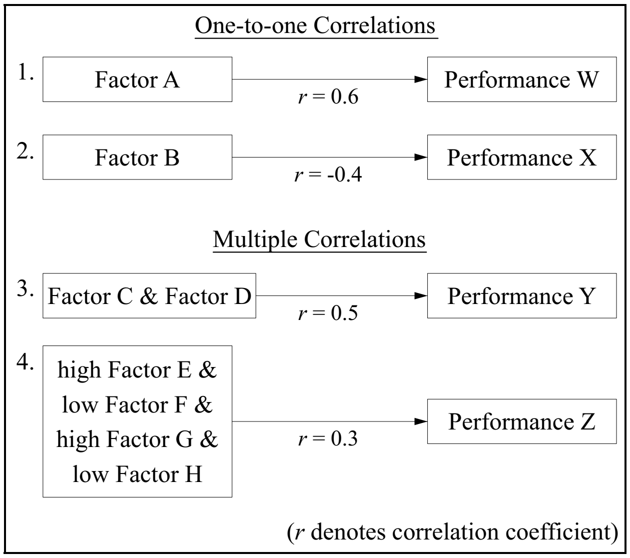

Figure 1. Examples of corrections between potential factors and performances.

An expert system [11-15] can be employed to solve this problem. A correlation between two factor means that there is a dependence or linkage between them (i.e. the magnitudes or values of the two factors changes depending on each other). Correlations can be positive or negative, strong or weak. Given a set of factors and a set of performances, determining the correlations between them is a problem of pattern recognition.

In this paper, the problem of finding multiple correlations between potential factors and performances is reduced to the regression problem in the field of machine learning. Support vector regression (SVR) is a technique used to build support vector machines (SVMs) for regression. The ability of SVR has been theoretically analyzed [1]. Given some known samples, an SVR model is built and correlation coefficients are defined. At first, personal attributes in each questionnaire are transformed into continuous-value scores to present the questionnaire instance. Then, questionnaire instances are used to build SVR prediction models. The target value of an instance can be predicted as the regression value. In our experiments, 12 questionnaires were collected for evaluating the performance of the developed SVR models. Correlations were found between a person's performance and personality characteristics.

Although SAT scores and high school academic records are good predictors of college grades, they cannot predict other important aspects of a student's performance, such as athletic ability, interests, and attitudes. Many psychology studies have found that personality is frequently related to many aspects of a student's academic performance. The Big Five personality model is thus adopted in this study to improve predictions [16].

The rest of the paper is organized as follows. The proposed prediction framework and the research design are presented in Section II. The experimental procedure is described and results are given in Section III. The conclusions are given in Section IV.

II. The SVR Model

Figure **2** shows the model of the proposed SVR prediction model. Inputs of the SVR model include three categories of factors, and outputs of the SVR model include three categories of performances. Each category contains multiple items, which are described below.

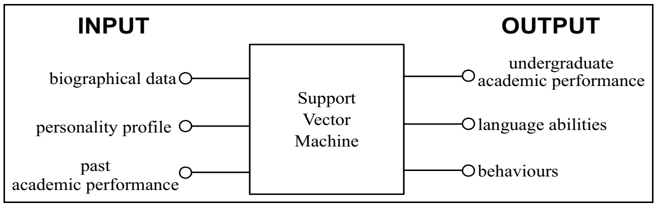

Figure 2. Proposed SVR prediction model

Input - factors

Biographical data includes 3 items – gender, birth order, and major, which are further divided into 9 subitems.

Personality profile includes 5 items – the five big-five personality traits (OCEAN).

The 5-point Likert scale items have choices: "strongly disagree" to "strongly agree", encoded to values - 0,1,2,3,4.
factors = { O, C, E, A, N }
Nx = number of items for measuring factor X.
Rx,i = Response of item i of factor X , R ∈ *{0,1,2,3,4}*
Formula to calculate the value of X, scaled to [0,1]:
Σ(Rx,i) / Nx / 4

Past academic performance includes 1 item – high school GPA (Grade Point Average).

Output - performances

Undergraduate academic performance includes 3 items – GPA, English courses, and Chinese courses.

Language ability includes 8 items – reading, listening, speaking, and writing skills of both English and Chinese languages.

Behaviours includes 4 items – Absence, Lateness, involvement of student clubs, and leadership.

Further descriptions of the items are given in the following. Of *biographical data*, *gender* indicates male and female, *birth order* indicates the order of birth among the siblings. The collective item *major* indicates the major taken in undergraduate study, which contains 9 items – *business*, *science*, *engineering*, *social science*, *humanities/art*, *law*, *education*, *design*, and *medicine* – to indicate which faculty the major belongs to. *Personality profile* indicates the levels of the five traits - *openness*, *conscientiousness*, *extraversion*, *agreeableness*, and *neuroticism* – adopted from big-five personality model. *High school GPA* has a single item indicating the GPA of the last year in high school.

Of *undergraduate academic performance*, *GPA* indicates the cumulative GPA in undergraduate years, while *English courses* and *Chinese courses* indicate the cumulative GPAs of related courses of Chinese and English languages respectively. Of *behaviours*, *absence* and *lateness* indicate the frequencies, *involvement of student clubs* indicates the activeness and involvement in student clubs and organisations, while leadership indicate the *leadership* skill and ability.

Typically, the objective of this paper is to find the potential correlations between the inputs (biographical data, personality profile, and past academic performance) and the outputs (undergraduate academic performance, language abilities, and behaviours) in Figure 3. SVR is applied for this problem since it has the ability of correlating certain features with the corresponding target value. Consequently, in the proposed model, one SVR model is generated for each item of output performances and all input factors. Then, this item can be predicted by the corresponding SVR model. It is expected to explore multiple correlations between input factors and each output item.

III. EXPERIMENT

The procedure of experiments is illustrated in figure 3. First, personal data are collected in a web questionnaire. Then, invalid data sets are filtered out by authors and the remaining data are used for experiments. At last, we use LIBSVM [19] to train SVR models by a part of the dataset, and remaining data are used for evaluating SVR prediction.

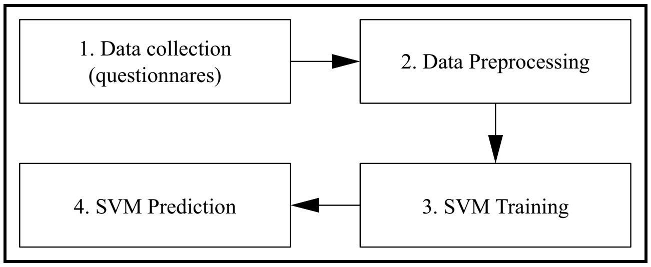

Figure 3 . Procedure of the experiment

A. Data set

In all 120 questionnaires, valid data, around 90 questionnaires, are randomly picked for training SVR models and others are used for evaluating their prediction. For building an SVR model to predict the specific item of output performances, each training instance should contain the input factors described in the proposed prediction model and the target output item. Notably, each training instance is generated by one certain questionnaire. Thus, the corresponding 15 SVR models are built according to items of output performances. Similarly, testing instances are generated by input factors and the target output item. A value of the target item is predicted by the corresponding SVR model and then is compared with its real value for analyzing the effectiveness of our model.

B. Prediction

TABLE I. EVALUATING THE ACCURACY OF THE PREDICTION

		Absent	Chinese	English	First_lng_listening_comprehe	First_lng_oral_comprehension	First_lng_reading_comprehens	First_lng_writing_comprehens	Involvement_of_student_clubs	Late	Leadership	Overal_average	Second_lng_listening_compreh	Second_lng_oral_comprehensio	Second_lng_reading_comprehen	Second_lng_writing_comprehen
	Std Dev	0.149	0.113	0.108	0.085	0.098	0.086	0.120	0.195	0.145	0.176	0.142	0.136	0.132	0.146	0.109
-s3-t0-c1	***MSE***	0.135	0.142	0.102	0.107	0.078	0.081	0.094	0.179	0.149	0.194	0.133	0.105	0.144	0.123	0.106
	StdDev-MSE	0.014	-0.029	0.005	-0.022	0.019	0.006	0.026	0.016	-0.004	-0.018	0.009	0.030	-0.012	0.023	0.003
-s3-t1-c1	***MSE***	0.123	0.115	0.088	0.087	0.063	0.081	0.093	0.188	0.129	0.198	0.141	0.089	0.128	0.115	0.121
	StdDev-MSE	0.026	-0.003	0.019	-0.003	0.035	0.005	0.027	0.007	0.016	-0.021	0.000	0.047	0.005	0.031	-0.012
-s3-t2-c1	***MSE***	0.126	0.125	0.090	0.097	0.069	0.080	0.091	0.174	0.135	0.191	0.132	0.096	0.143	0.123	0.108
	StdDev-MSE	0.023	-0.013	0.017	-0.012	0.028	0.006	0.029	0.021	0.010	-0.015	0.009	0.039	-0.011	0.023	0.001
-s3-t0-c99999	***MSE***	0.134	0.153	0.094	0.105	0.099	0.089	0.110	0.178	0.169	0.192	0.125	0.113	0.141	0.125	0.120
	StdDev-MSE	0.015	-0.040	0.014	-0.021	-0.002	-0.003	0.010	0.017	-0.024	-0.016	0.017	0.022	-0.009	0.021	-0.010
-s3-t1-c99999	***MSE***	0.202	0.186	0.165	0.247	0.097	0.087	0.297	0.272	0.265	0.345	0.310	0.273	0.223	0.219	0.158
	StdDev-MSE	-0.053	-0.073	-0.057	-0.162	0.001	0.000	-0.177	-0.077	-0.120	-0.169	-0.168	-0.138	-0.091	-0.073	-0.049
-s3-t2-c99999	***MSE***	0.245	0.219	0.125	0.214	0.090	0.080	0.320	0.411	0.224	0.305	0.290	0.294	0.267	0.320	0.138
	StdDev-MSE	-0.096	-0.107	-0.018	-0.130	0.007	0.007	-0.200	-0.216	-0.079	-0.129	-0.149	-0.158	-0.134	-0.174	-0.029

Investigation of the accuracy of the prediction, rather than through arithmetic mean of differences better actual values and predicted values, can be done through comparing standard deviation and mean square error. MSE-Std Dev should has a value larger than zero to yield a good prediction accuracy.

Among the prediction results of the three sets of SVR parameters and fifteen prediction items, 44% of which have MSE-Std Dev values of greater than 0.01, meaning that SVR is able to predict these values rather than guessed by chance. Among the fifteen items, the SVR did the prediction the best in *absent*, *Chinese speaking*, *Chinese writing*, *English listening*, and *English reading*. Among the results with the three sets of SVR parameters, the prediction accuracies are close.

IV. Conclusion

A comprehensive framework for predicting students' performance in university based on SVR and the Big Five personality model was proposed. The system gives students suggestions on their potential abilities. The results show that the problem of finding multiple correlations between potential personal factors and performances can be solved using the proposed SVR model.

The participants in the experiment were undergraduates and people who had finished school years ago. The latter group might not have correctly remembered their past grades and past performance, which could hurt the validity of the prediction. This problem will be addressed in a future study.

References

[1] V. Vapnik. The Nature of Statistical Learning Theory, Springer, N. Y., 1995. ISBN0-387-94559-8.

[2] Burton, N.W. and Ramist, L., Predicting success in college: SAT studies of classes graduating since 1980, College Board Research Report, vol. 2, 2001.

[3] Bridgeman, B. and Pollack, J. and Burton, N., Predicting grades in different types of college courses, College Board Research Report, vol. 1, 2008.

[4] Cooper, C.I., Predicting Persistence of College Freshmen Using Neural Networks, Technological Developments in Education and Automation, pp. 145-148, 2010.

[5] Wongkhamdi, T. and Seresangtakul, P., A Comparison of Classical Discriminant Analysis and Artificial Neural Networks in Predicting Student Graduation Outcomes, Proceedings of the Second International Conference on Knowledge and Smart Technologies, 2010.

[6] Lykourentzou, I. and Giannoukos, I. and Mpardis, G. and Nikolopoulos, V. and Loumos, V., Early and dynamic student achievement prediction in e-learning courses using neural networks, Journal of the American Society for Information Science and Technology, Vol. 60, No. 2, pp. 372-380, 2009.

[7] Lye, C.T. and Ng, L.N. and Hassan, M.D. and Goh, W.W. and Law, C.Y. and Ismail, N., Predicting Pre-university Student's Mathematics Achievement, Procedia-Social and Behavioral Sciences, vol. 8 , pp. 299-306, 2010.

[8] Thai-Nghe, N. and Drumond, L. and Krohn-Grimberghe, A. and Schmidt-Thieme, L., Recommender system for predicting student performance, Procedia Computer Science, Vol. 1, No. 2, pp. 2811-2819, 2010.

[9] Fang, N. and Lu, J., Work in progress-a decision tree approach to predicting student performance in a high-enrollment, high-impact, and core engineering course, Frontiers in Education Conference, 2009. FIE'09. 39th IEEE, pp. 1-3, 2009.

[10] Romero, C. and Espejo, P.G. and Zafra, A. and Romero, J.R. and Ventura, S., Web usage mining for predicting final marks of students that use Moodle courses, Computer Applications in Engineering Education, 2010.

[11] Anderson, J.L., Predicting Final GPA of Graduate School Students: Comparing Artificial Neural Networking and Simultaneous Multiple Regression., College and University, Vol. 81, No. 4, pp. 11, 2006.

[12] Naik, B. and Ragothaman, S., Using neural networks to predict MBA student success, College Student Journal, Vol. 38, No. 1, pp. 143-149, 2004.

[13] Cripps, A., Using artificial neural nets to predict academic performance, Proceedings of the 1996 ACM symposium on Applied Computing, pp. 33-37, 1996.

[14] Cooper, C., A Neural Network-Based Decision Support System for Identifying and Remediating At-Risk Students, The International Journal of Applied Management and Technology, pp. 255, 2007.

[15] Rusli, N.M. and Ibrahim, Z. and Janor, R.M., Predicting students' academic achievement: Comparison between logistic regression, artificial neural network, and Neuro-fuzzy, Information Technology, 2008. ITSim 2008. International Symposium on, vol. 1, pp. 1-6, 2008.

[16] Trapmann, S. and Hell, B. and Hirn, J.O.W. and Schuler, H., Meta-analysis of the relationship between the big five and academic success at university, Journal of Psychology, Vol. 215, No. 2, pp. 132-151, 2007.

[17] Agresti, A. and Finlay, B., Statistical methods for the social sciences, 1997.

[18] Landry, R. and Amara, N. and Lamari, M., Utilization of social science research knowledge in Canada, Research policy, Vol. 30, No. 2, pp. 333-349, 2001.

[19] C.-C. Chang and C.-J. Lin. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011.