Key attribute for predicting student academic performance

Conference Paper · October 2018

DOI: 10.1145/3290511.3290576

1 author:

Sachio Hirokawa
Kyushu University
300 PUBLICATIONS **1,201** CITATIONS

SEE PROFILE

Key Attribute for Predicting Student Academic Performance

Sachio Hirokawa
Research Institute for Information Technology, Kyushu University
Motooka 744, Fukuoka 819-0395, JAPAN
hirokawa@cc.kyushu-u.ac.jp

ABSTRACT

Predicting student final score from student's attributes is an important issue of learning analytic. Not only to achieve high prediction performance but also to identifying the key attributes is an important research theme. This paper evaluated exhaustively the prediction performance based on all possible combinations of four types of attributes -- behavioral features, demographic features, academic background, and parent participation. The behavioral features are given as numerical data. But, we represented them as pair of an attribute name and the value. This vectorization yields 417 dimensional data, while naively represented data has 68 dimension. By applyig support vector machine and feature selection, we obtained the optimal prediction performance, with respect to feature selection, with accuracy 0.8096 and F-measure 0.7726. We confirmed that the behavioral feature is so crucial that the accuracy reaches 0.7905 without other features except behavioral feature. The combination of behavior feature and demographic feature gained F-measure 0.7662.

CCS Concepts

•Applied computing→Education→E-learning

•Theory and algorithms for application domains→Machine learning→Support vector machine

Keywords

Learning analytic Machine Learning; Support Vector Machine; .Feature selection

1. INTRODUCTION

Due to digitization of the educational environment, information on learning activities as well as basic information on learners is accumulated. Research on Learning Analytics (LA) and Educational Data Mining (EDM) are exected for improving education by using these information. It is expected to estimate the final grade of the student from demographic features, behavioral features or characteristics of subjects.

If the characteristics of students with good grades can be clarified, we will be able to incorporate such feature for the lectures. Individual guidance becomes possible if the characteristics of students with poor grades are known in advance. Extraction of keys factors that affect student's final scores and estimation of scores using them are important issues of LA. Even if a model with high predition performance is obtained, if it is a black box whose structure is unknown, we cannot use for education. In other words, an interpretable model is required.

Table 1 Top 10 Feature with respect to scr*freq

rank	scr*freq	scr	freq	type	attribute:value
1	288.1730	0.9971	289	b	absance:Under-7
2	137.1089	0.6960	197	p	relation:Mum
3	83.5525	0.3095	270	p	parentAns:Yes
4	70.5187	0.4030	175	d	gender:F
5	59.9354	0.2550	235	a	semester:S
6	41.6952	0.2836	147	d	grade:G-02
7	33.8592	0.1701	199	d	stage:lowerlevel
8	30.7095	0.1716	179	d	nationality:KW
9	27.9629	0.9321	30	a	topic:Biology
10	24.2239	1.1535	21	a	topic:Math

Table 1 shows the top ten features when ranking is given by the product of the score of the attributes and the frequency of the attributes. They are the positive feature of positive data, i.e., students who received H (hight) score as their final evaluation. The top feature “absance:Unider-7” measns that the absence number is seven or less. The second feature “relation:Mum” means that Mum is responsible for education. The abbriviated types "b", "p", "d", and "a" in the fifth column represent four kinds of attributes, behavioral features, parent participation, demographic features, and academic background.

Table 2 Top 10 Features with respect to scr

rank	scr	freq	Type	attribute:value
1	1.4834	5	b	discussion:81
2	1.3608	6	b	reaisedhands:100
3	1.3596	1	b	discussion:73
4	1.1645	6	b	visited:17
5	1.1535	21	a	topic:Math
6	1.1286	12	b	visited:70
7	1.1207	3	b	reaisedhands:98
8	1.1020	7	b	view:6
9	1.1020	4	b	discussion:2
10	1.0982	6	b	view:51

Table 2 shows the top ten features with respect to the feature selection in descending order of scores. The top feature “discussion:81” means that the student obtained 81 point as his or her evaluation of discussion. In Table 1, various types of features

appeare, but in Table 2, the behavioral feature is the majority. We used the score of the feature as in Table 2, instead of the product of the score and the frequency as in Table 1, since the former yields the best prediction performance than the later. From this, we can see that behavioral feature plays an important role in predicting student final scores. [Amerieh 2016] points out that behavioral feature is important based on mutual information. In addition, they compared the prediction performance between when using behavioral features and when not using them. They showed that using behavioral feature has higher prediction performance. In this paper, we evaluate how much influence each attribute type has on the prediction performance. Specifically, the prediction performance was evaluated for 15 kinds of all of the possible combinations of the four types of features.

2. Dataset and Student Attribute

In this paper we analyze the information of 480 students provided by [Amerieh 2016]. Students have the final scores ranked by H (Hight), M (Middle) and L (Low). We use this score as our prediction target. Table 3 shows the input variables to predict the final score. They are classified in four types of attributes: demographical features, academic background, parent participation, and behavioral features. Behavioral features 9, 10, 11, and 12 represent the evaluation of students' behavior as a numerical value in the range of 0 to 100. All other 12 attributes are categorical data.。

Table 3 Student Attributes

	abbr	type	attribute	value	sample
0	d	demographic	gender	categorical	M
1	d	demographic	NationalITy	categorical	KW
2	d	demographic	PlaceofBirth	categorical	KuwaIT
3	d	demographic	StageID	categorical	lowerlevel
4	d	demographic	GradeID	categorical	G-04
5	a	academic	SectionID	categorical	A
6	a	academic	Topic	categorical	IT
7	a	academic	Semester	categorical	F
8	p	parent	Relation	categorical	Father
9	b	behavioral	raisedhands	number	15
10	b	behavioral	VisITedResources	number	16
11	b	behavioral	AnnouncementsView	number	2
12	b	behavioral	Discussion	number	20
13	p	parent	ParentAnsweringSurvey	categorical	Yes
14	p	parent	ParentschoolSatisfaction	categorical	Good
15	p	behavioral	StudentAbsenceDays	categorical	Under-7

3. Vectorization of Dataset

Prior research [Amrieh2016, Chohan2018] that analysed the same dataset do not describe vectorization of dataset in detail. It seems that they followed the standard method of handling a set of item names and values as one attribute. In other words, it seems that the instance of each item was vectored as one attribute. The 15th attribute "StudentAbsenceDays" is a behavioral feature. But, it is a categorical feature. They seem to be handled in the same way. The dimension corresponding to such categorical features total 68 dimensions. It seems that four types of behavioral features other than “StudentAbsenceDays” are vectorized as one real value. Therefore, dataset was expressed in 71 dimensions. In [Amrieh2016], they predicted the final score of H, M and L with only ten attribues 10,15,9,11,13,1,8,2,12 and 14 based on the feature selection with respect to the the mutual information (MI).

In this paper, we use the categorical representation [Mobasher2017] for behavioral features given as numerical data in the rage of 0 to 100. Table 1 shows the values of a student for each of the 16 types of attributes. We consider a numerical data as a categorical data and consider a pair of attribute name and value as a word, and each student is vectorized as a document containing those words. Thus, this student is represented as a BOW (bag of words) that contain the following 16 words: 0:M, 1:KW, 2:KuwaIT, 3:lowerlevel, 4:G-04, 5:A, 6:IT, 7:F, 8:Father,9:15, 10:16, 11:2, 12:20, 13:Yes, 14:Good, and 15:Under-7.

Behavioral features are numbers ranging from 0 to 100, but not all values appear in the dataset. In fact, there were only 82 values for "Raised Hands" (No. 9), 89 values for "visited" (No. 10), 88 values for "view" (No.11) and 90 values for "discussion" (No.12). Thus, the behavioral features of numerical data are represented as 349 dimensional space of the pairs of attribute name and boolean values. The whole dataset is represented as 417 dimensional boolean space by merging other categorical features.

4. Prediction of Hight Performance Students

In this paper, we evaluate which combination of the four types of attributes is effective for prediction. We firstly represeted four kinds of attributes by initial letters "a" (academic background), "b" (behavioral features), "d" (demographic features) and "p" (parent information). Secondly, we conducted exhaustive experiment to evaluate the prediction performance with respect to all possible 15 combinations of the four attributes. We used SVM and feature selection [Amrieh2016,Lesinski2016] as a machine learning method.

In this chapter, we describe the procedures and results in detail about the prediction of students whose evaluation is H (Hight). This method consists of two steps. In the first step, we apply SVM to construct a model for 480 students vectorized using all attributes. The component of the vector is represented as a pair "w_i:v_i" of attribute number (w_i) from 0 to 15 and Boolean values (v_i) for that attribute. As mentioned in the previous section, the dimension of this vector is 714. Therefore, in the constructed model, the predicted value *guess* (u_k) of each student u_k is calculated as the inner product of a the vectors of the weight *scr* (w_i:v_i) and the occurrence vectors δ (w_i:v_i,u_k), where δ (w_i:v_i,u_k) is 1 if w_i:v_i

appears in u_k, 0 o.w. In this paper, we call the score *scr(wi:vi)* as SVM-score of the component w_i:v_i.

$$\text{guess}(u_k) = \beta + \sum_{i=1}^{M} scr(w_i : v_i) * \delta(w_i, u_k)$$

In the second step, we select the total of 2 * N attributes of the top N and the bottom N components with respect to the SVM-score *scr(wi:vi)*. Those 2*N attributes are used for vectorization of students. We search the optimum N that yields the best prediction performance by changing N. In [Amrieh2016], they limited to ten attributes with high values with respect to mutual information. In [Adachi2016], it is shown that the feature selection by SVM-score *scr(wi:vi)* has higher prediction performance than feature selection by mutual information. We conducted the performance evaluation by 10-fold cross validation both in the first and in the second stage.

Table 4 Prediction Performance of H students withouth Feature Selection

	precision	recall	F−measure	accuracy
abdp	0.7804	0.3060	0.4341	0.7743
bdp	0.8352	0.2797	0.4050	0.7700
bp	0.7717	0.2502	0.3694	0.7526
abp	0.7319	0.2853	0.3963	0.7525
adp	0.7917	0.1835	0.2782	0.7380
abd	0.6872	0.2043	0.2973	0.7364
bd	0.6848	0.1781	0.2745	0.7345
dp	0.6833	0.1179	0.1923	0.6590
ab	0.6305	0.1878	0.2748	0.6588
b	0.6117	0.1457	0.2289	0.6461
ap	0.6278	0.0941	0.1575	0.6385
p	0.4076	0.2759	0.2955	0.5172
d	0.4000	0.0373	0.0667	0.3631
ad	0.1667	0.0216	0.0377	0.2306
a	0.0000	0.0000	0.0000	0.0000

Table 5 Prediction Performance of H students with Feature Selection

	N	precision	recall	F−measure	accuracy
abdp	90	0.7484	0.8190	0.7794	0.8718
abp	90	0.8018	0.7707	0.7766	0.8716
bp	60	0.7452	0.8396	0.7850	0.8676
bd	100	0.7207	0.7592	0.7354	0.8398
bdp	80	0.6529	0.8963	0.7513	0.8260
ab	90	0.6252	0.8487	0.7091	0.8064
b	40	0.9090	0.3287	0.4772	0.7883
abd	60	0.6000	0.8978	0.7058	0.7863
ap	8	0.6055	0.5226	0.5505	0.7594
adp	20	0.6582	0.4451	0.5185	0.7532
d	10	0.4331	0.1760	0.2481	0.7012
ad	8	0.5142	0.3643	0.3941	0.6879
dp	20	0.7000	0.1839	0.2757	0.6681
p	0	0.4076	0.2759	0.2955	0.5172
a	4	0.2013	0.1314	0.1210	0.4416

Table 4 shows the prediction performance of high score students (H) without feature selection. Table 5 shows the prediction performance with feagure selection. This first columnf of the Table 4 and 5 shows the combination of attributes. Figures 1, 2 and 3 show the prediction performance by only one kind of attributes, the prediction performance by two kinds of attributes, and the prediction performance by three or more types of attributes. These figures show the change in accuracy when N is changed. From Figure 1, we can see that "b" (behavioral feature) has the highest performance as one type of attribute and attains the best accuray 0.78 around N=40. The next highest prediction performance as a single attribute is "d" (demographic features), with N = 10 and accuracy of 0.70. Accuracies for "a" and "p" are at least 0.50. For the two kinds of attributes combination, only three of "bp", "bd" and "ab" that contains "b" are over 0.80. Since it is not greatly improved from the performance when b alone, the importance of b isconfirmed. With three or more attribute combinations, "adp" that does not include "b" is the lowest, and performance is lower than "bp", "bd", "ab", and "ap" that use only two types. Here also, it is clear that the effect of "b" is large.

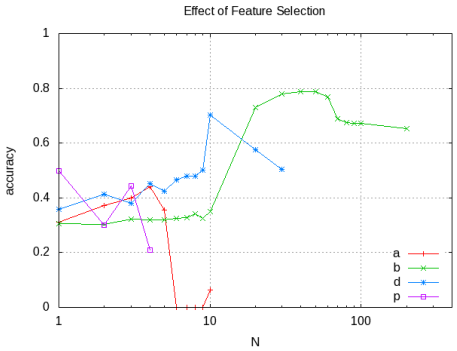

Figure 1 Accuracy with Single Attribute

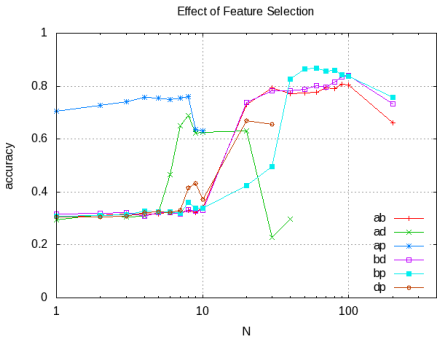

Figure 2 Accuracy with Attribute Pair

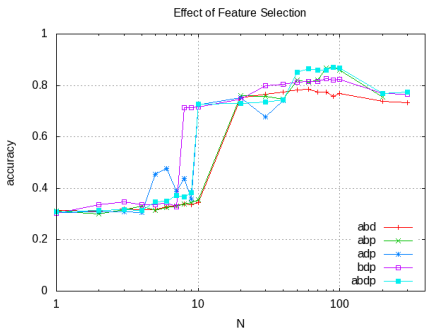

Figure 3 Accuracy with Triple & All Attributes

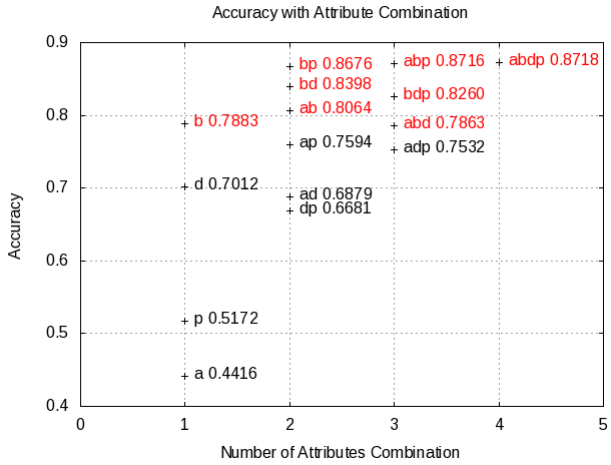

Figure 4 Accuracy with Attribute Combination for H

Figure 4 shows the prediction performance of each attribute combination in two dimensions. The horizontal axis indicates the number of combinations of attributes and the vertical axis indicates accuracy. Note that the attribute combinations containing “b” are drawn in red color. Only "b" and "d" are more than 0.70 in one kind of attribute. No combination of two or more achieves better accuracy over 0.80 without "b". It is whorthwhile to notice that the combination "adp", which is the complement of "b", achieves 0.75 and that "b" achieves the similar performance 0.78. In other words, it can be said that about 75% of the prediction can be achieved based only with behavioral features or withouth behavioral features. Thus, behavioral features is necessary, in order to obtain further prediction performance.

5. Attribute Combination

5.1 Optimal Feature Selection for H, M and L

We might think, in general, that the more the attributes used for prediction, the higher the prediction performance. In this paper, we applied feature selection of [Adachi2017, Sakai2012] for all combinations of fourtypes of attributes "a", "b", "d" and "p", and searched for the number N of feature selection that yields the optimum accuracy. Tables 5, 6, and 7 show the prediction performance by optimal feature selection when students with grades H, M, L are taken as positive examples.

For "abdp" using all the attributes, the accuracy of H, M, L and H are 0.8718, 0.6015 and 0.9488, respectively. We can understand that the prediction of "M" students is harder than that of "H" and "L". However, as shown in the second column of Table 6, the accuracy by "b" for M is 0.7156, which is higher than when using all attributes. Also in Table 5 and 7, the accuracy of "b" alone is clearly higher than that of other single attributes "a", "d" and "p". Furthermore, the performance by "b" is higher than that of the combination of two attributes "ad" and "dp". Another interesting observation the three kinds of combinations is that in any of Table 5, 6, 7, "adp" is clearly lower in performance than the other three combinations. These can be interpreted as the importance of "b".

Table 6 Optimal Feature Selection for M

M	N	precision	recall	F-measure	accuracy
a	4	0.5066	0.5721	0.5327	0.5725
b	80	0.7081	0.6330	0.6600	0.7156
d	10	0.4224	0.5635	0.4723	0.4761
p	1	0.0375	0.0643	0.0474	0.0556
ab	80	0.5867	0.8477	0.6883	0.6705
ad	20	0.5107	0.5219	0.5140	0.5748
ap	5	0.6068	0.2302	0.3219	0.5937
bd	90	0.5803	0.8546	0.6869	0.6664
bp	200	0.7514	0.1697	0.2734	0.6087
dp	20	0.5121	0.2555	0.3313	0.5680
abd	70	0.5678	0.8363	0.6741	0.6498
abp	50	0.5454	0.8930	0.6745	0.6222
adp	40	0.5967	0.0973	0.1655	0.5865
bdp	300	0.7048	0.1887	0.2900	0.6173
abdp	200	0.6783	0.2011	0.3059	0.6015

Table 7 Optimal Feature Selection for L

L	N	precision	recall	F-measure	accuracy
a	3	0.2565	0.6265	0.3503	0.4276
b	30	0.9333	0.5330	0.6694	0.8677
d	10	0.4100	0.1769	0.1803	0.6970
p	0	0.2437	0.2265	0.2286	0.3807
ab	100	0.8245	0.9090	0.8611	0.9286
ad	9	0.3246	0.8343	0.4549	0.4853
ap	7	0.5936	0.6628	0.6159	0.7951
bd	60	0.8396	0.9252	0.8762	0.9295
bp	100	0.8630	0.8613	0.8531	0.9267
dp	6	0.4141	0.8815	0.5547	0.6292
abd	100	0.8435	0.9084	0.8714	0.9288
abp	70	0.8512	0.8895	0.8667	0.9349
adp	20	0.6175	0.5782	0.5771	0.7840
bdp	70	0.9142	0.9035	0.9064	0.9535
abdp	80	0.8879	0.9343	0.9016	0.9488

H	N	precision	recall	F-measure	accuracy
a	4	0.2013	0.1314	0.1210	0.4416
b	40	0.9090	0.3287	0.4772	0.7883
d	10	0.4331	0.1760	0.2481	0.7012
p	0	0.4076	0.2759	0.2955	0.5172
ab	90	0.6252	0.8487	0.7091	0.8064
ad	8	0.5142	0.3643	0.3941	0.6879
ap	8	0.6055	0.5226	0.5505	0.7594
bd	100	0.7207	0.7592	0.7354	0.8398
bp	60	0.7452	0.8396	0.7850	0.8676
dp	20	0.7000	0.1839	0.2757	0.6681
abd	60	0.6000	0.8978	0.7058	0.7863
abp	90	0.8018	0.7707	0.7766	0.8716
adp	20	0.6582	0.4451	0.5185	0.7532
bdp	80	0.6529	0.8963	0.7513	0.8260
abdp	90	0.7484	0.8190	0.7794	0.8718

5.2 Average Prediction Performance

Table 8 shows the macro average of accuracies for feature selection that maximizes accuracy for H, M and L. Figures 5 and 6 show the macro average of accuracies and F-measures, respectively. We used macro average in this pape as overall evaluation measure, since thethe previous research used macro average The accuracy is improved by feature selection in any attribute combination. Also, pair reaches higher performance than single attribute. There is not much difference between pair and triple.

In particular, regarding attribute combination including "b", all of single, pair and triple combinations attain almost 80%. Looking at Figure 6 on F-measure, the difference in performance due to attribute combination is more significant than in Figure 5. In particular, we can be see that the performance of combination without "b" is low.

Table 8 Macro Average of Optimal Feature Selection

average	precision	recall	F-measure	accuracy
a	0.3215	0.4433	0.3347	0.4806
b	0.8501	0.4982	0.6022	0.7905
d	0.4218	0.3055	0.3002	0.6248
p	0.2296	0.1889	0.1905	0.3178
ab	0.6788	0.8685	0.7528	0.8018
ad	0.4498	0.5735	0.4543	0.5827
ap	0.6020	0.4719	0.4961	0.7161
bd	0.7135	0.8463	0.7662	0.8119
bp	0.7865	0.6235	0.6372	0.8010
dp	0.5421	0.4403	0.3872	0.6218
abd	0.6704	**0.8808**	0.7504	0.7883
abp	0.7328	0.8511	**0.7726**	**0.8096**
adp	0.6241	0.3735	0.4204	0.7079
bdp	0.7573	0.6628	0.6492	0.7989
abdp	**0.7715**	0.6515	0.6623	0.8074

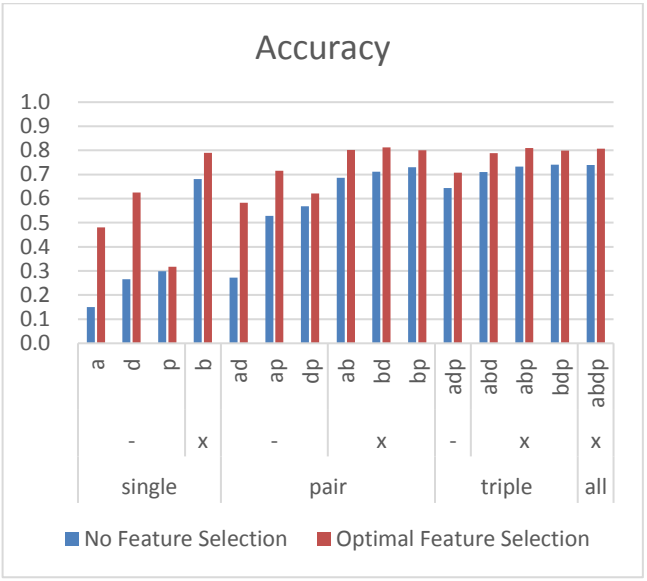

Figure 5 Macro Accuracy

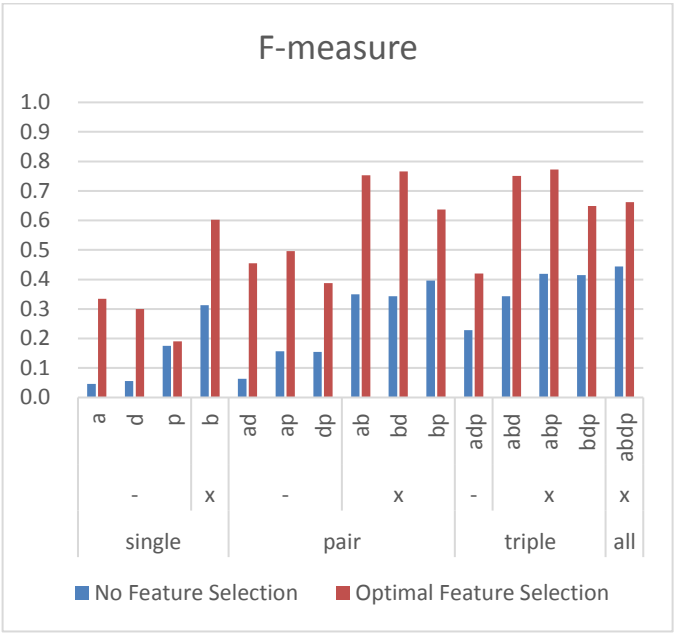

Figure 6 Macro F-measure

6. Related Work

The data we analysed in the present paper was first analysed in [Amerieh2016]. It is considered as one of benchmarks of learning analytic. Infact, [Kapur2017, Rahman2017, Tanabe2017] analysed the same data with different method, where [Rahman2017] showed the best accuracy 84.3 percent as far as the author know.

To predict student's academic performance from students' feature is the most important issue in learning analytic. However, it alone is not a goal. To understand the reason of student's performance is much more important issue. [Cortez2008] is one of the early

research in this direction, where they showed that student performance is influenced by past evaluations. But they found other relevant features, e.g., number of absences, parent's job and eductaion. [Mobasher2016] obtained good prediction performance with REP treee. They used the model to construct the rules descirbed with demographic data, study related and psychological caracteristics. They proposed to use the rules as for recommending advice students, teachers and parents. [Lesinski 2016] suceeded to predict graduation of students with 95% accuracy using input varialbes of academic data , demographic data and other indicators. But they did not invesitigate the most important feature. [Koutina2011] predicted the final grade of the postgraduate students from demographic feature, in-term performance (1st, 2nd and 3rd writing assignments) and presence in class. They found that presence in class is one of the most imporant feature. The confirmed that student occupation , type of bachelor degree do not improve accuracy. The present paper would be the first approach that evaluated precisely and exhaustively the influence of attribute combination.

7. Conclusion and Further Work

In this paper, for 480 students, we conducted the exhaustive evaluation of the prediction performance of students of levels H, M and L based on 15 kinds of attribute combinations. There are four types of attributes -- demographic, academic, parent and behavioral features. Three of the behavioral attributes are numerical data of the scores in the range of 0 to 100 points. All other attributes are categorical data. Each student was regarded as a document that contains pairs of attribute name and the attribute value and are represented as 417 dimensional data. Applying SVM and feature selection [Sakai2012], high prediction performance was obtained with F-measre of 66% and accuracy of 80%. Earlier research [Amerieh 2016] also said the importance of behavioral feature. In this paper, we evaluate the predition performances for all combinations (15 ways) of the four kinds of attributes and show the effectiveness of the behavioral attribute.

In this paper, we used only the simplest SVM as a machine learning method. [Rahman 2017] achieved an accuracy of 84.3% with ensembler filtering, which is better than that of the present paper. Other vectorization and other machine learning methods will be evaluated as further work.

We confirmed that the behavior feature is the key to predict students' academic performance. However, it does not imply that other features are meaning less. In fact, the accuracy with the demographic feature is very close to that with the behavior feature. Further research is necessary to analyze the relatioship among the four attributes.

References

[1] Adachi Y., Onimura N., Yamashita T. , Hirokawa S., (2017) Classification of Imbalanced Documents by Feature Selection, Proc. ICCDA 2017, pp.228-232

[2] Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016), Mining Educational Data to Predict Student's academic Performance using Ensemble Methods, International Journal of Database Theory and Application, 9(8), pp. 119-136

[3] Kapur, B., Ahluwalia, N., Sathyaraj, R. (2017), Comparative Study on Marks Prediction using Data Mining and Classification Algorithms, International Journal of Advanced Research in Computer Science, Vol. 8 Issue 3, pp. 632-636

[4] Koutina, M., Kermanidis, K.L. (2011), Predicting postgraduate students' performance using machine learning techniques, IFIP Advances in Information and Communication Technology 364 AICT(PART 2), pp. 159-168

[5] Lesinski, G., Corns, S., Dagli, C. (2016) , Application of an Artificial Neural Network to Predict Graduation Success at the United States Military Academy,Procedia Computer Science 95, pp. 375-382

[6] Mobasher, G., Shawish, A., Ibrahim, O. (2017) , Educational data mining rule based recommender systems, CSEDU 2017 - Proceedings of the 9th International Conference on Computer Supported Education 1, pp. 292-299

[7] Sakai T. , Hirokawa S., Feature Words that Classify Problem Sentence in Scientific Article, Proc. iiWAS2012, pp.360-367, 2012

[8] Tanabe, T., Kagari, K., Kitanaka Y., Kazuhiro, T., Hirokawa S. (2017), Finding Key Integer Values in Many Features for Learner's Academic Performance Predication, Proc. ICETC2017, pp.167-171