Discretization method to optimize logistic regression on classification of student's cognitive domain

by Puput Wanarti

Submission date: 01-Jul-2019 12:14PM (UTC+0700)
Submission ID: 1148333540
File name: Puput_matecconf_aasec2018_yuni.pdf (474.36K)
Word count: 3148
Character count: 15740

Discretization method to optimize logistic regression on classification of student's cognitive domain

Yuni Yamasari[1,3*], *Puput W.* Rusimamto[4], *Naim* Rochmawati[3], *Dwi F.* Suyatno[3], *Setya C.* Wibawa[3], *Supeno M. S.* Nugroho[1,2], and *Mauridhi H.* Purnomo[1,2]

[1]Institut Teknologi Sepuluh Nopember, Department of Electrical Engineering, Surabaya, Indonesia
[2]Institut Teknologi Sepuluh Nopember, Department of Computer Engineering, Surabaya, Indonesia
[3]Universitas Negeri Surabaya, Department of Informatics, Faculty of Engineering, Surabaya, Indonesia
[4]Universitas Negeri Surabaya, Department of Electrical Engineering, Faculty of Engineering, Surabaya, Indonesia

Abstract. The accuracy level of the student determination in a class often has been paid less attention in educational data mining. So, this paper studies how to improve the performance of classification method to reach the higher of level accuracy. Therefore, we optimize logistic regression using equal frequency discretization method. Here, we test the student data by three intervals, four intervals, and five intervals. For logistic regression, we implement two regularization types, namely: lasso, ridge. Furthermore, to evaluate the results, we use the random sampling technique. Additionally, we measure the results by four classifier metrics, namely: F1, precision, accuracy, and recall. The experimental result shows that this method can be applied to optimize the logistic regression. On logistic regression_lasso and logistic regression_ridge, the three intervals achieve the highest of accuracy level. They can improve the accuracy level about 9% - 9.4%, respectively.

1 Introduction

Nowadays, research in education area focuses on processes enhancement to make the better of the education environment. Research workes many efforts to achieve this goal. One of the examples is the implementation of educational data mining [1]. There are many tasks in data mining, for examples: clustering [2], classification, analysis association, etc. [3].

Specifically, researchers focus on the classification of educational data. Classification using four methods: Decision Tree, Rule Induction, Naïve Bayes and Neural Network in research [4] is addressed to predict the performance of student academic. Also, a decision tree is also explored to predict the effective learning in research [5]. Guo et al. also predict the student performance using the classification method. Here, they exploit deep learning [6]. Next, the classification is done to predict the student dropout factor by using Induction Rule and Decision Tree in [7]. Different from the others, research focuses on the improvement of the Particle Swarm Classification to classify the question level in an examination [8]. Additionally, L.Ge et al. in [9] apply SVM as a classifier to predict about extraversion and introversion traits on the student. Almost previous research is addressed for prediction. As far as we know, classification methods need dataset which has a label because of classification as the supervised learning.

In contrary, the others research generates information about predicting using Linear Regression. In [10], the research build model to forecast the grade level of student achievement which assists the teacher in the early handle to the poor student. Another research applies linear regression to predict students performance in final examination [11]. Additionally, the performance predicting is also studied by [12]. Here, to achieve the goal, research employ patterns of student activity that one of bag classifier is a linear regression. Next, the automated marker quality is offered by the applied of this method in research [13]. Lastly, the prediction is also done by the research [14]. It is addressed to predict the student's psychomotor domain. Nevertheless, all research does not study on about the performance enhancement. Also, linear regression is usually applied to data in the form of continuous variables.

Therefore, our research explores logistic regression as classification method to handle the categorical variables. Furthermore, our research also improves the accuracy level of the classifier performance to generate the most valid information. Finally, this information is useful to decide how many classes in the data labeling.

2 Methods

In this section, we illustrate the proposed framework consisting of many steps as follows:

* Corresponding author: yuniyamasari@unesa.ac.id

Step 1: Extracting features based on category.

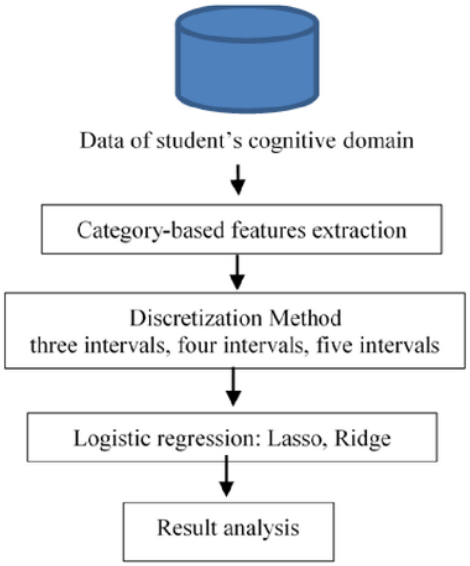

Fig. 1. The proposed framework.

In this step, we extract features of student's cognitive domain to improve the performance of educational data mining [15]. This step produces ten features, namely: number of main items which are answered by the student (MID), the percentage of the right answer and all answer of main items (MID%_true), time elapsed to answer the main items (Time_MID), student score of main items (Score_MID). Next, the number of guidance/scaffolding items which are answered by the student (GID), the percentage of the right answer and all answer to guidance/scaffolding (GID%_true). Then, the time elapsed to answer the guidance/scaffolding (Time_GID), student score of guidance/scaffolding (Score_GID), number of the accessed hints (Hint) and the sum of MID and GID (Total_score).

Step 2: Doing discretization method

Student data are a continues variable which has been extracted is discretized to many intervals: three intervals, four intervals, and five intervals. Here, we use discretization method called equal frequency. This method is addressed to optimize the classification process, so the classification performance is better than before. Furthermore, this information can be used to make the best decision about how many classes of our data set.

Step 3: Applying logistic regression

We propose logistic regression as a classification method. This step is done to know how many intervals which can produce the most optimal of the classification process. A logistic regression learns a logistic regression model from data. So, logistic regression learning algorithm is as a learner.

Logistic regression is a regression model with categorical dependent variable [16]. It is a simple understanding way of finding the β parameters on equations:

$$y = \begin{cases} 1 & \beta_0 + \beta_1 x + \varepsilon > 0 \\ 0 & else \end{cases} \quad (1)$$

The standard logistic distribution spreads an error symbolized. Particularly, in machine learning algorith logistic regression is an important algorithm used to model the probability of a random variable Y is or 1 given experimental data. Additionally, logistic regression evaluates the relationship between the categorical dependent variable and one or more independent variables by estimating babilities using a logistic function having the formula as follows:

$$\sigma(t) = \frac{e^t}{e^t + 1} \quad \frac{1}{1 + e^t} \quad (2)$$

Table 1. Performance of Logistic regression for Discretization-three-intervals.

Repeat Train/Test	Train set size:	Lasso				Ridge			
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
2	10%	0.562	0.708	0.573	0.926	0.642	0.724	0.594	0.926
	20%	0.788	0.835	0.815	0.855	0.821	0.848	0.841	0.855
	30%	0.801	0.829	0.807	0.852	0.824	0.868	0.885	0.825
	40%	0.839	0.891	0.891	0.891	0.839	0.882	0.872	0.891
	50%	0.796	0.838	0.861	0.816	0.827	0.827	0.85	0.895
	60%	0.859	0.875	0.824	0.933	0.846	0.871	0.844	0.9
3	10%	0.61	0.744	0.643	0.882	0.705	0.765	0.669	0.892
	20%	0.769	0.843	0.848	0.839	0.791	0.848	0.857	0.839
	30%	0.799	0.832	0.838	0.827	0.828	0.873	0.896	0.852
	40%	0.831	0.884	0.884	0.884	0.831	0.884	0.884	0.884
	50%	0.81	0.867	0.875	0.86	0.816	0.881	0.852	0.912
	60%	0.838	0.86	0.833	0.889	0.821	0.867	0.867	0.867
Average		0.775	0.834	0.808	0.871	0.799	0.845	0.8259	0.8782

Table 2. Performance of Logistic regression for Discretization-four-intervals.

Repeat Train/Test	Train set size:	Lasso				Ridge			
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
2	10%	0.506	0.923	0.857	1	0.562	0.915	0.915	1
	20%	0.673	0.941	0.889	1	0.641	0.97	0.941	1
	30%	0.706	0.955	0.913	1	0.743	0.966	0.933	1
	40%	0.72	0.935	0.878	1	0.703	0.96	0.923	1
	50%	0.735	0.952	0.909	1	0.755	0.952	0.909	1
	60%	0.756	0.923	0.857	1	0.731	0.96	0.923	1
3	10%	0.553	0.931	0.871	1	0.602	0.959	0.92	1
	20%	0.675	0.954	0.911	1	0.65	0.966	0.935	1
	30%	0.73	0.969	0.94	1	0.716	0.969	0.94	1
	40%	0.672	0.913	0.871	1	0.667	0.956	0.915	1
	50%	0.741	0.957	0.918	1	0.741	0.957	0.918	1
	60%	0.744	0.947	0.9	1	1	0.96	0.923	1
Average		0.684	0.942	0.893	1	0.709	0.958	0.9246	1

Table 3. Performance of Logistic regression for Discretization-five intervals.

Repeat Train/Test	Train set size:	Lasso				Ridge			
		Acc	F1	Prec	Rec	Acc	F1	Prec	Rec
2	10%	0.517	0.806	0.707	0.935	0.562	0.857	0.803	0.919
	20%	0.517	0.806	0.707	0.935	0.562	0.857	0.803	0.919
	30%	0.676	0.907	0.898	0.917	0.669	0.893	0.836	0.958
	40%	0.712	0.938	0.974	0.905	0.763	0.927	0.95	0.905
	50%	0.745	0.899	0.886	0.912	0.796	0.886	0.861	0.912
	60%	0.846	0.915	0.871	0.964	0.833	0.947	0.931	0.964
3	10%	0.523	0.822	0.761	0.892	0.572	0.878	0.835	0.925
	20%	0.632	0.914	0.914	0.914	0.671	0.904	0.882	0.926
	30%	0.691	0.923	0.93	0.917	0.716	0.914	0.873	0.958
	40%	0.695	0.942	0.983	0.905	0.751	0.935	0.951	0.921
	50%	0.776	0.879	0.839	0.922	0.816	0.95	0.987	0.922
	60%	0.846	0.92	0.889	0.952	0.821	0.941	0.93	0.952
Average		0.681	0.889	0.863	0.923	0.711	0.907	0.887	0.932

Here, we assume that t is a linear function of the single explanatory variable x. So, t is expressed as follows:

$$t = \beta_0 \quad \beta_1 x \qquad (3)$$

Consequently, the logistic function can be written as:

$$p(x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} \quad \text{a} \qquad (4)$$

This method is used for predicting a dependent variable that is categorical as in the previous step.

Additionally, in computer science, especially in machine learning field, regularization is addressed to prevent overfitting or to solve an ill-posed, problem with introducing additional information. Here we use two regularizations: lasso [17] and ridge [17-19] to solve it.

Step 4: Analysing results

In this step, we analyze the experimental results. We compare each other to find the best result which is indicated by the optimal value for every metrics relating to the classification performance.

3 Result and Discussion

The execution of the proposed framework is described in this section. Next, we analyze the experimental results. For the first result, we visualize applying to the student data which is discretized to three intervals, four intervals and 5-intervals on logistic regression. Here, we set a parameter on logistic regression using lasso and ridge. The first result is presented in Figure 2-7. In these Figures, we can show that Figure 2-3, Figure 4-5, and Figure 6-7 are obtained from discretization-three

intervals, discretization-four intervals, and discretization-five intervals, respectively.

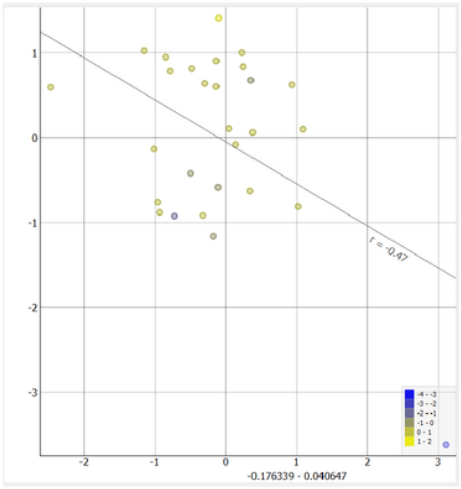

Fig. 2. Discretization-three intervals-lasso

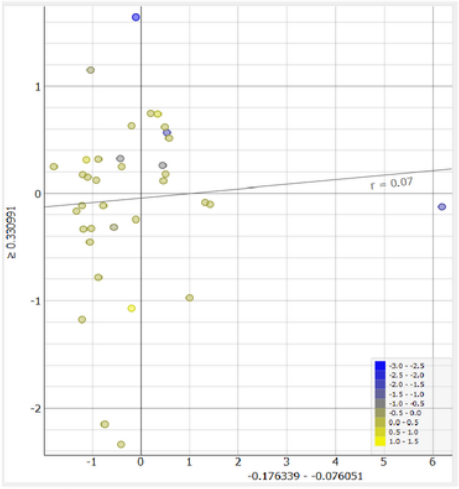

Fig. 4. Discretization-four intervals-lasso

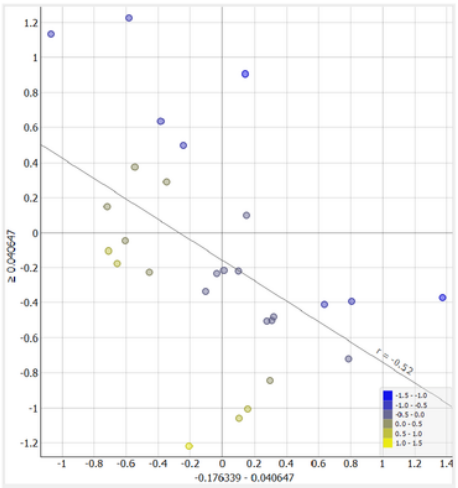

Fig. 3. Discretization-three intervals-ridge

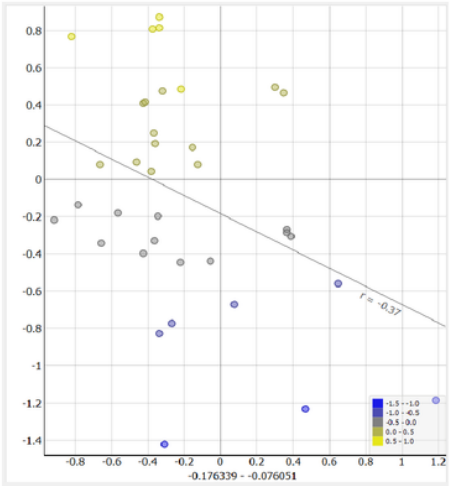

Fig. 5. Discretization-four intervals-ridge

Furthermore, we explore them based on parameters of logistic regression. Here, we know that parameter ridge has the higher of correlation value (r) than lasso for all intervals of discretization method. On lasso, value r of three intervals, four intervals and five intervals are 0.47, 0.07 and 0.23, respectively. For ridge, 0.52, 0.37 and 0.30 are achieved by three intervals, four intervals and five intervals, respectively. Furthermore, the highest of correlation is achieved by discretization- three intervals on all parameter of logistic regression.

Relating to the performance of logistic regression as a classifier, we employ many metrics, namely: accuracy (Acc), F1, precision (Prec), and recall (Rec). Table 1-3 shows the results of running for all intervals and all parameters on logistic regression. We evaluate the classifier model using the percentage split technique.

In addition, data training uses many sizes: 10%, 20%, 30%, 40%, 50% and 60% and iteration of training/testing: 2-3. Table 2 shows the performance of discretization method with five intervals applied on logistic regression for lasso and ridge regularization.

For lasso, the highest of accuracy level is about 0.846 that is achieved on train set size 60%. Additionally, the highest of F1, precision, and recall are approximately 0.942, 0.983 and 0.964, respectively. Next, ridge reaches the highest of accuracy, F1, precision and recall around 0.833, 0.95, 0.987 and 0.964, respectively.

The experimental result of discretization method with four intervals is showed in Table 3. Here, lasso attains the highest of accuracy, F1, precision, and recall are about 0.756, 0.969, 0.94 and 1, respectively. Then, the highest of accuracy, F1, precision, and recall around 1,

0.97, 0.941 and 1, respectively. For discretization method with three intervals, the experimental result is showed in Table 3. Here, the highest of accuracy, F1, precision, and recall are 0.859, 0.891, 0.891 and 0.933, respectively. These metrics are achieved by Lasso.

Meanwhile, on the ridge, the highest of accuracy, F1, precision, and recall on the ridge are approximately 0.846, 0.884, 0.896 and 0.926, respectively.

Further, when all of the experimental results are compared each other, the highest of accuracy level is achieved by three intervals on both regularization lasso and ridge about 0.775 and 0.799, severally. This achievement is based on its average.

Here, we can infer that discretization method with three intervals can increase 9% - 9.4% than the other intervals.

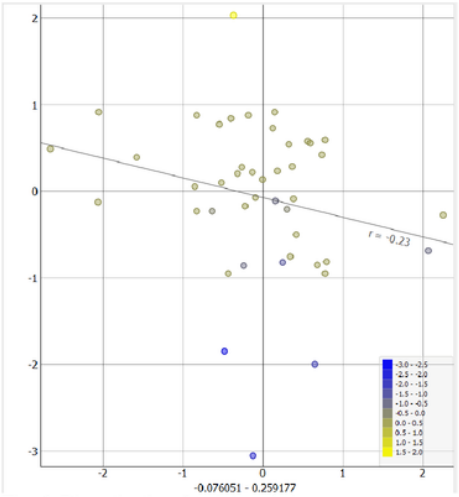

Fig. 6. Discretization-five intervals-lasso

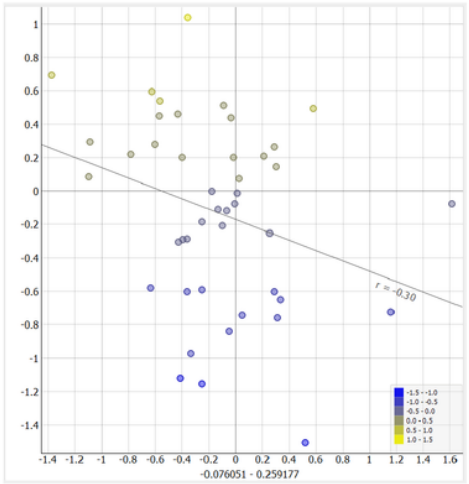

Fig. 7. Discretization-five intervals-ridge

4 Conclusion

In this paper, we propose the discretization method to optimize lgiogistic regression on the educational dataset. Also, we apply two regularizations, namely: lasso and ridge. Here, discretization method with three intervals is able to enhance the performance of classification on student's cognitive domain for all regularizations of logistic regression.

The authors would like to thank Universitas Negeri Surabaya for its support through the funding program of the International Conference.

References

1. L. C. Liñán, Á. Alejandro, and J. Pérez, "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution Learning Analytics: Intelligent Decision Support Systems for Learning Environments," RUSC. Univ. Knowl. Soc. J., vol. **12**, no. 3, pp. 98–112, 2015.
2. Y. Yamasari, S. M. S. Nugroho, R. Harimurti, and M. H. Purnomo, "Improving the cluster validity on student's psychomotor domain using feature selection," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 460–465.
3. P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to data mining*. Pearson Addison Wesley, 2005.
4. R. Asif, A. Merceron, and M. K. Pathan, "Predicting Student Academic Performance at Degree Level: A Case Study," I.J. Intell. Syst. Appl. Intell. Syst. Appl., vol. **1**, no. 1, pp. 49–61, 2015.
5. N. A. Shukor, Z. Tasir, and H. Van der Meijden, "An Examination of Online Learning Effectiveness Using Data Mining," in *Procedia - Social and Behavioral Sciences*, 2015, vol. **172**, pp. 555–562.
6. B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," in *International Symposium on Educational Technology, ISET 2015*, 2016.
7. A. Pradeep, S. Das, and J. J. Kizhekkethottam, "Students dropout factor prediction using EDM techniques," in *Proceedings of the IEEE International Conference on Soft-Computing and Network Security, ICSNS 2015*, 2015.
8. A. A. Yahya, "Swarm intelligence-based approach for educational data classification," *J. King Saud Univ. – Comput. Inf. Sci.*, 2017.
9. L. Ge, H. Tang, Q. Zhou, Y. Tang, and J. Lang, "Classification Algorithms to Predict Students' Extraversion-Introversion Traits," in *2016 International Conference on Cyberworlds (CW)*, 2016, pp. 135–138.
10. K. Kongsakun, "An improved recommendation model using linear regression and clustering for a

private university in Thailand," in *2013 International Conference on Machine Learning and Cybernetics*, 2013, pp. 1625–1630.

11. F. Widyahastuti and V. U. Tjhin, "Predicting students performance in final examination using linear regression and multilayer perceptron," in *2017 10th International Conference on Human System Interactions (HSI)*, 2017, pp. 188–192.
12. K. Casey and D. Azcona, "Utilizing student activity patterns to predict performance," Int. J. Educ. Technol. High. Educ., vol. **14**, no. 1, p. 4, Dec. 2017.
13. F. Grivokostopoulou, I. Perikos, and I. Hatzilygeroudis, "An Educational System for Learning Search Algorithms and Automatically Assessing Student Performance," Int. J. Artif. Intell. Educ., vol. **27**, no. 1, pp. 207–240, Mar. 2017.
14. R. Harimurti, Y. Yamasari, Ekohariadi, Munoto, and B. I. G. P. Asto, "Predicting student's psychomotor domain on the vocational senior high school using linear regression," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 448–453.
15. Y. Yamasari, S. M. S. Nugroho, I. N. Sukajaya, and M. H. Purnomo, "Features extraction to improve performance of clustering process on student achievement," in *2016 International Computer Science and Engineering Conference (ICSEC)*, 2016, pp. 1–5.
16. D. Freedman, *Statistical models : theory and practice*. Cambridge University Press, 2009.
17. R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58. WileyRoyal Statistical Society, pp. 267–288, 1996.
18. A. E. Hoerl, R. W. Kennard, and R. W. Hoerl, "Practical Use of Ridge Regression: A Challenge Met," Appl. Stat., vol. **34**, no. 2, p. 114, 1985.
19. N. R. Draper and R. C. van Nostrand, "Ridge Regression and James-Stein Estimation: Review and Comments," Technometrics, vol. **21**, no. 4, p. 451, Nov. 1979.

Discretization method to optimize logistic regression on classification of student's cognitive domain

ORIGINALITY REPORT

15%	11%	7%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.matec-conferences.org Internet Source	4%
2	Submitted to School of Business and Management ITB Student Paper	3%
3	en.wikipedia.org Internet Source	2%
4	repository.ubaya.ac.id Internet Source	1%
5	Submitted to Istanbul Aehir Aniversitesi Student Paper	1%
6	"Advances in Information Retrieval", Springer Science and Business Media LLC, 2019 Publication	1%
7	"Neural Information Processing", Springer Nature, 2017 Publication	1%
8	Submitted to Staffordshire University	

	Student Paper	1%
9	Submitted to Sikkim Manipal University Student Paper	<1%
10	amysartgallery.com Internet Source	<1%
11	Submitted to University of Cumbria Student Paper	<1%
12	Submitted to Higher Education Commission Pakistan Student Paper	<1%
13	Submitted to Jawaharlal Nehru University (JNU) Student Paper	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		