IIAI Open Conference Publication Series
IIAI Letters on Institutional Research
Vol. 001, LIR018
DOI: https://doi.org/10.52731/lir.v001.018

Evaluation of Predictive Models in Institutional Research Based on Multi-Objective Optimization

Nobuhiko Kondo [*], Toshiharu Hatanaka [†],
Takeshi Matsuda [*]

Abstract

In institutional research (IR), the use of predictive models based on machine learning has attracted significant attention, especially for predicting students at risk of dropping out (at-risk students) and academic success. Since various evaluation metrics for predictive models in IR can be considered, the tradeoffs among them must be taken into account in model selection. Thus, this study considers the model selection process, as a multi-objective optimization problem, and proposes a framework to visualize the results of evaluating model candidates by multiple evaluation indicators, which are important in the IR context. Specific examples of numerical experiments using actual data are also presented to summarize its effectiveness and challenges.

Keywords: Institutional research, Predictive modeling, Machine learning, Multi-objective optimization, Decision-making

1 Introduction

In recent years, institutional research (IR), which is responsible for evidence-based decision-making support, has become increasingly important in Japanese higher education, due to the growing demand for educational quality assurance and accountability to stakeholders. In Europe and the United States, systematic operation of learning analytics, in which advanced data analysis of education and learning is used to support education and learning, has been actively studied. In this regard, the use of predictive models based on machine learning has attracted attention as a promising method [1].

Traditionally, IR has used visualization and statistical methods to explain the data, but the extension and sophistication of IR function through the use of prediction has been widely studied in recent years by integrating it with learning analytics. In general, when building a predictive model, data is divided into training and test data, and the model is trained such that the evaluation of the test data is high under a certain evaluation metric. However, there are various evaluation indicators for this model. While models with high predictive accuracy, e.g., in terms of correct response rates and prediction errors, are useful, models with as few explanatory variables as possible can also be considered desirable, especially in terms of the explainability of the model. Moreover, there are several indicators of prediction performance, each of which includes a different meaning.

[*] Tokyo Metropolitan University, Tokyo, Japan
[†] The University of Fukuchiyama, Kyoto, Japan

For effective use of predictive models in IR, it would be desirable to have a framework that evaluates candidate models by using various indicators, and visualizes the results of the evaluation so that decision-makers can consider appropriate ones. Therefore, this study considers the model selection process, as a multi-objective optimization problem, and proposes a method that visualizes a set of predictive models under various algorithms and parameters, which are important in the IR context. It also discusses the applicability of the proposed method and future issues by presenting specific examples of numerical experiments using actual educational data.

2 Predictive Models in IR and Their Evaluation

2.1 Predictive Models in IR

In the field of learning analytics, there are numerous examples of predicting at-risk students such as those dropping out of classes, etc. Early-alert systems for early detection, warning, and intervention of at-risk students are also being employed at many educational institutions [2]. Meanwhile, it is possible to predict high-performers in a similar framework to that for at-risk students.

In IR for education and learning, predictive models are generally used to predict future performance based on certain variables related to students, and to provide some type of support/intervention based on the results. In Japan, there have been many studies on the prediction of withdrawal from school, and it is expected that the prediction of academic success, which can be achieved within a similar framework, will become an important theme in the future.

One of the distinctive challenges of IR is that it is difficult to design a generic predictive model that can be used for any institution. This is because the background and properties of the data, as well as the quality and quantity of the available variables, can widely vary from institution to institution. Hence, a framework for model-building should be developed that allows each institution to individually build and examine predictive models according to its context.

2.2 Indicators for Evaluating Predictive Models in IR

There are two commonly used evaluation indicators for predictive models: precision and recall, both of which are used in classification problems. When the number of positive examples correctly predicted as positive is denoted as TP, the number of negative examples incorrectly predicted as positive is denoted as FP, and the number of positive examples incorrectly predicted as negative is denoted as FN, precision is defined as TP/(TP + FP) and recall is defined as TP/(TP + FN). Precision also indicates the percentage of correct output results from the predictive model, while recall indicates the percentage of correctly detecting what we want to detect, This can be thought of as corresponding to type I and type II errors in statistical inference, respectively. Additionally, these indicators are in a tradeoff relationship, and the balance between them varies, depending on the threshold of discrimination.

Although the tradeoff between precision and recall generally appears in classification problems, IR-specific considerations should also be made in this regard. Assuming the actual use of predictive models in IR, it is likely that numerous applications will involve screening students that need support based on the prediction results, or conducting an automatic intervention. Since one of the important issues is how to detect students who actually require assistance without any omissions, it will be necessary to obtain a higher recall rate. Conversely, an increasing recall will most likely lead to a decrease in precision, in which case many of the true targets can be found, but the predictions are frequently missed. Such failure to predict can lead to increased costs and it may be undesirable from an ethical perspective, considering the selection of targets for support. From these viewpoints, it is preferable to have as high precision as possible. Meanwhile, the acceptable degree of precision and recall for an organization depends on the organization's policies, costs, and other restrictive conditions. It also depends on the organization's situation as to how much weight to place on precision and recall, respectively. Although it is possible to make an overall judgment based on the F-measure value, which is the harmonic mean of precision and recall, it is more important to make a judgment based on the actual values of precision and recall.

In addition, since the variables related to students handled in IR may be in the order of several hundred to several thousand or more, it is desirable that the interpretation of explanatory variables be as easy as possible when considering the use of predictive models for student support. From the viewpoint of the explainability of the model, it would also be useful to consider the number of explanatory variables the model includes, as an indicator for evaluating the goodness of the model. While the aforementioned evaluation indicators are representative, other evaluation indicators may be considered appropriate for each institution's situation.

2.3 Multi-objective Optimization of Predictive Models in IR

Multiple evaluation indicators, as described in section II.B, often have a tradeoff relationship with one another. In this study, we propose a framework to visualize the tradeoffs of model evaluations by various indicators, in order to enable the selection of appropriate predictive models.

When there are m evaluation indicators to be considered and the corresponding evaluation functions for each of them are f_1, f_2, …, f_m, the model selection can be regarded as a multi-objective optimization problem [3]. The multi-objective optimization problem is as follows:

$$\begin{array}{lll} \text{Minimize} & f_m(\boldsymbol{x}), & m = 1, \dots, M; \\ \text{subject to} & g_j(\boldsymbol{x}) \geq 0, & j = 1, \dots, J; \\ & h_k(\boldsymbol{x}) = 0, & k = 1, \dots, K; \\ & x_i^{(L)} \leq x_i \leq x_i^{(U)}, & i = 1, \dots, d. \end{array}$$

Here, $\boldsymbol{x} = (x_1, x_2, \dots, x_d)^T$ is a d-dimensional vector of the decision variable, while $x_i^{(L)}$ and $x_i^{(U)}$ are the lower and upper bounds in the decision space, respectively.

In the multi-objective optimization problem, the concept of dominance is used to consider a tradeoff among evaluation functions. x_1 is said to dominate x_2 if:

$$\begin{array}{lll} & \forall i = 1,2, \dots, M & f_i(x_1) \leq f_i(x_2) \\ \text{and} & \exists j = 1,2, \dots, M & f_j(x_1) < f_j(x_2). \end{array}$$

The solutions that are not dominated by any other solutions are called non-dominated solutions (Fig. 1). In general, many non-dominated solutions exist. A set of non-dominated solutions can be found because it is impossible to simultaneously optimize all of the evaluation functions in multi-objective optimization problems.

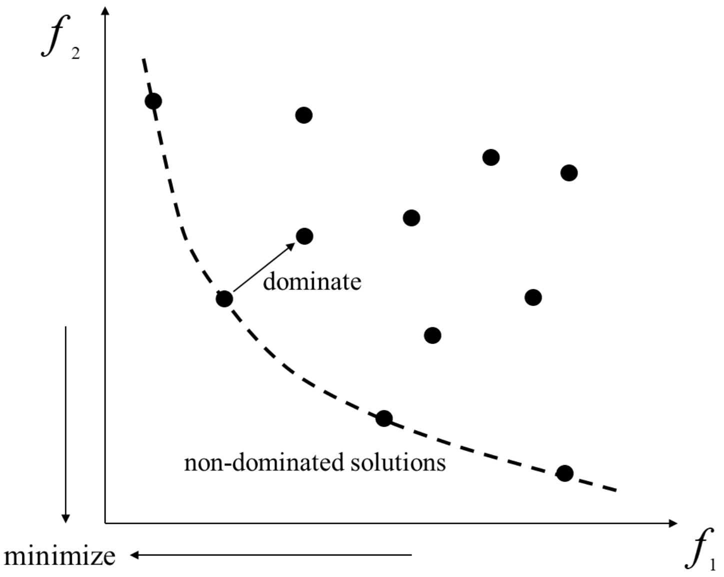

Figure 1: Domination in multi-objective optimization problems. (This figure is based on the figure in [4].)

We have previously proposed a multi-objective optimization for predictive models in IR, using a multi-objective genetic algorithm for variable selection in a predictive model of academic success[4]. In this study, we consider the following procedure as a generalized framework:

1) Train predictive models under various algorithms and parameters, and generate multiple candidates of predictive models.

2) Evaluate each candidate model using m evaluation indicators, and obtain m evaluation values for each model.

3) Calculate the dominance relationship among all of the candidate models based on the m evaluation values, and obtain a set of models as non-dominated solutions.

4) Visualize the tradeoffs among the evaluation indicators for the set of non-dominated models.

5) Based on the visualized tradeoffs, examine the adopted model.

Since this framework is independent of the type of evaluation indicator, it is possible to use any indicator that the educational institution considers important. The next section presents some examples of this procedure by means of numerical experiments using actual educational data.

3 Numerical Experiments

3.1 Outline of One Numerical Experiment

In this experiment, we considered the problem of predicting the status of students at the beginning of their fourth year (based on data up to the end of their first year) by using the data of students enrolled in the 2015–2017 academic year at University A. Here, two types of prediction problems were considered: (i) predicting academic success; and (ii) predicting at-risk students. The academic success prediction in (i) determines whether the student has a high cumulative grade point average (GPA), while the at-risk prediction in (ii) determines whether the student will take a leave of absence or withdraw from the university. In the former case, the students with a GPA (between 0 and 4) of 3.0 or higher were considered to be successful and defined as positive cases, while in the latter case, the students who had taken a leave of absence or dropped out of school were defined as positive cases.

Since it is likely that in actual IR operation, a model learned from students' data in one entrance year is often used to make predictions for students in subsequent entrance years, the training data and test data were split by entrance year in this study. Specifically, the data for the students enrolled in the 2015–2016 school year was used as the training data (N = 3275) and the data for the students enrolled in the 2017 school year was used as the test data (N = 1628). The number of positive examples was 798 (24.4%) for the training data and 365 (22.4%) for the test data in the problem (i), and 104 (3.2%) for the training data and 70 (4.3%) for the test data in the problem (ii). This data was obtained with permission in accordance with the rules and procedures regulated by University A for the use of educational data in academic research.

A total of 43 variables were used as explanatory variables, including: entrance examination category, affiliation, gender, English test (three items, twice, at the time of admission and at the end of the first year, respectively), GPA, number of credits earned, credit acquisition rate (three types: first semester, second semester, and total in the first year, respectively), and GPA by subject category in the first year (25 types).

This experiment was performed in Python 3.7.6 using the scikit-learn package.

3.2 Case of Two Evaluation Indicators

Numerical experiments were conducted based on the procedures in sections II.C 1) to 4). For the problems (i) and (ii), we first trained several predictive models with different learning algorithms and various parameter settings. The algorithms used were logistic regression and random forest.

For logistic regression, L1 and L2 regularization were used, and in both cases, the models were trained by varying the regularization parameter C with values of 0.001, 0.01, 0.1, 1, and 10. For random forest, the models were trained by varying the number of decision trees as 10, 20, and 40. For both algorithms, the threshold for binary classification for the output value of the objective variable, which takes values in [0, 1], was varied in increments of 0.1 in the range [0.1, 0.9]. Hence, $(2 \times 5 + 3) \times 9 = 117$ candidate predictive models were generated.

Finally, we considered precision and recall as the evaluation indicators of the model and obtained the set of non-dominated solutions. As a result, 19 non-dominated solutions were obtained for the problem (i) and 17 were obtained for the problem (ii). The evaluation results of the candidate models for problems (i) and (ii) are shown in Figures 2–3. Based on these visualizations, decision-makers can select the model to be used, by considering the risks and benefits.

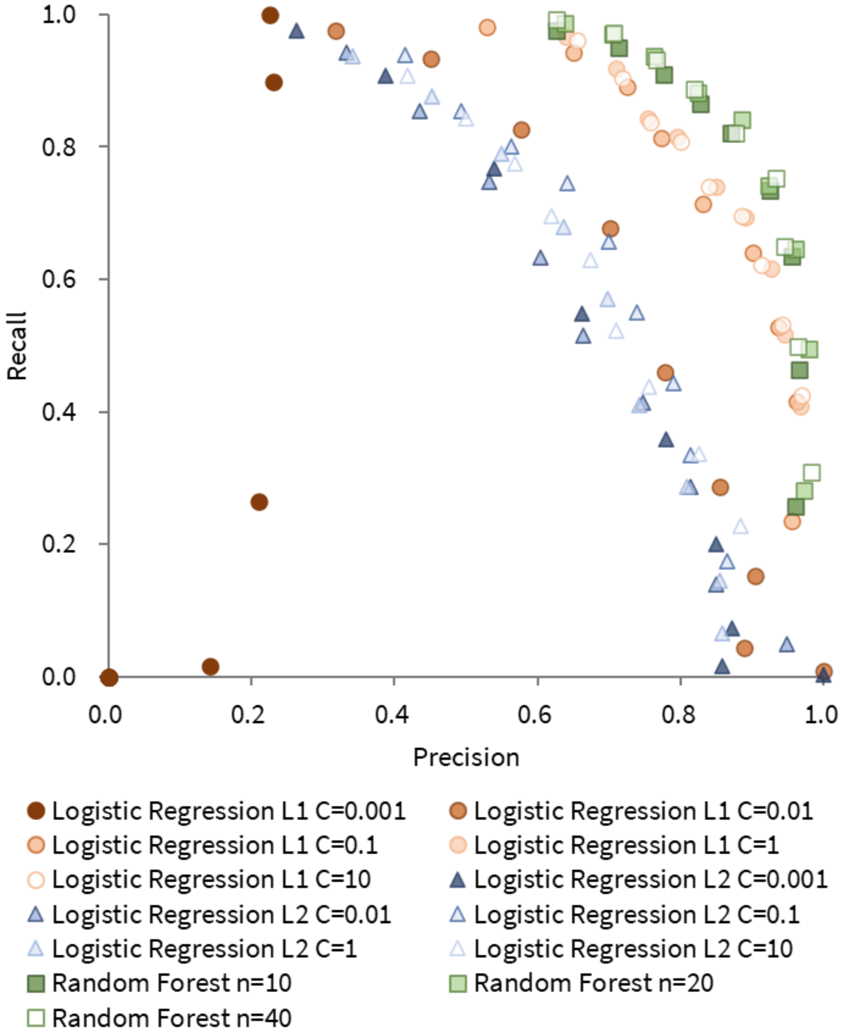

Figure 2: Candidate predictive models in the problem (i).

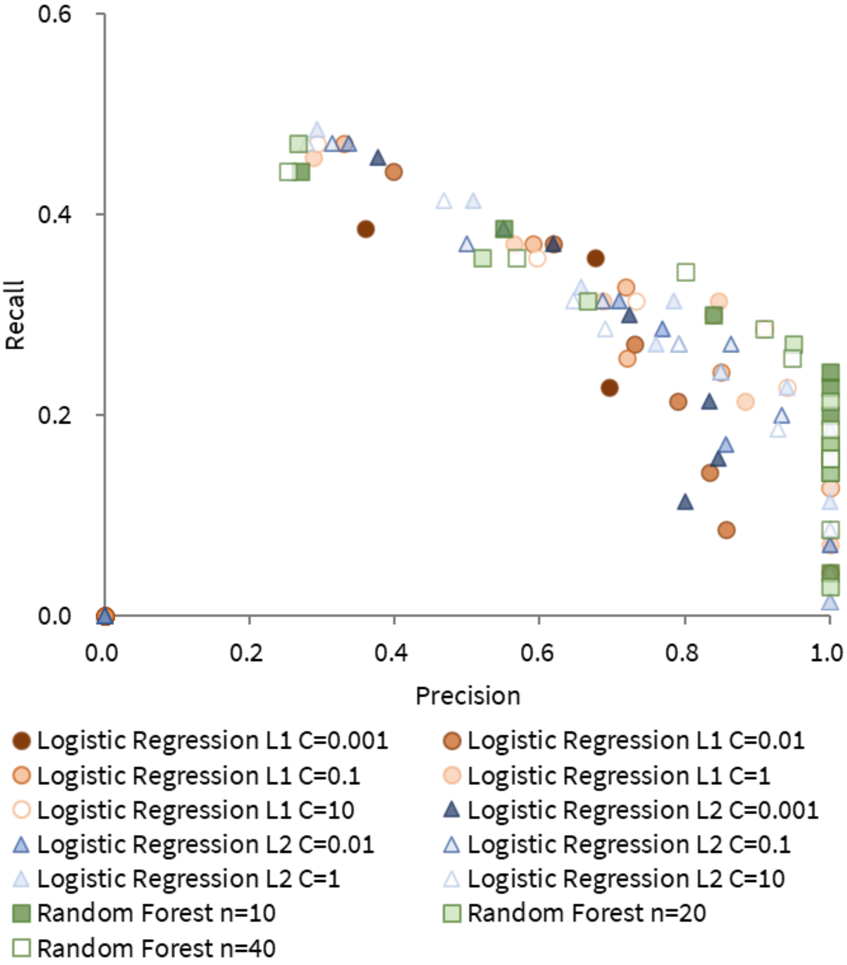

Figure 3: Candidate predictive models in the problem (ii).

3.3 Case of Three Evaluation Indicators

As an example of assuming three evaluation indicators, in addition to precision and recall, we considered the number of explanatory variables as the third evaluation indicator, as one that evaluates the model’s explainability. In this case, by using the method of variable selection, the variables that contribute to the prediction are selected from among all of the candidate explanatory variables. Hence, the fewer the number of variables, the better the model is. Among the various methods of variable selection, we used L1 regularization, which can make the coefficients of some explanatory variables zero. Thus, these variables are removed from the model.

In order to illustrate the three evaluation indicators in two dimensions, the number of explanatory variables is represented by a gradation according to the magnitude of their values and by a side note of the values themselves in the two axis figures of precision and recall. Figure 4 presents the problem (i), while Figure 5 shows the problem (ii). The figures indicate that some solutions that are dominated by a certain solution with respect to precision and recall are conversely dominated (fewer in number) in terms of explanatory variables. This type of visualization allows us to examine the appropriate model while checking the tradeoffs among the three evaluation indicators.

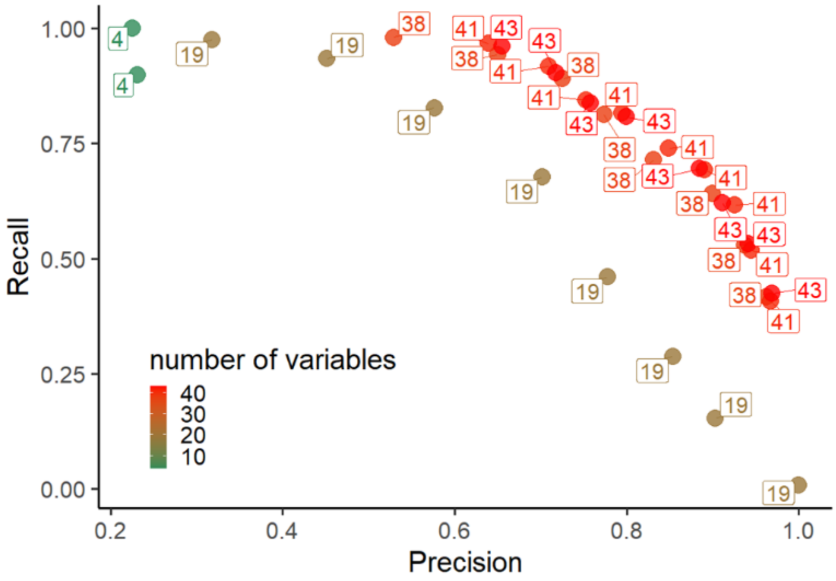

Figure 4: Visualization of the three evaluation values of the candidate models by L1 regularization (problem (i))

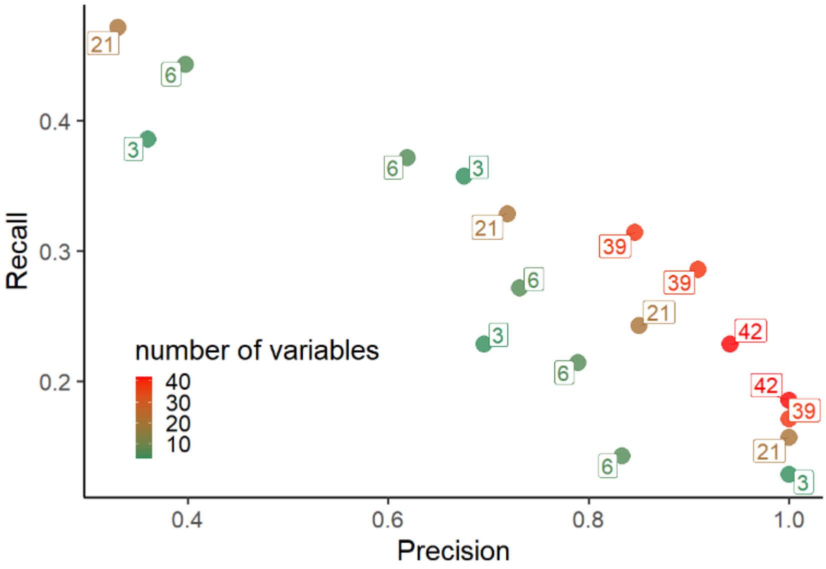

Figure 5: Visualization of the three evaluation values of the candidate models by L1 regularization (problem (ii))

3.4 Future Works

As described earlier, by generating multiple candidate predictive models and visualizing the tradeoffs related to multiple possible evaluation indicators in IR, it is possible to select the appropriate one according to the organization's situation, which can lead to more appropriate decision-making.

In this experiment, we generated and compared the models for all combinations because we limited the number of algorithms, the hyperparameters of the algorithms, and the step settings of the threshold values in advance. However, if a wider range of possibilities is explored by increasing the number of search algorithms or by refining the search granularity of parameters, then a combinatorial explosion will occur, making the search for a solution more difficult. Therefore, the use of meta-heuristics for multi-objective optimization, such as multi-objective genetic algorithms, should be considered.

4 Conclusion

This study described a method to evaluate and visualize predictive models in IR, from the perspective of multi-objective optimization, and enable the consideration of appropriate model selection according to the organizational context. This was based on the results of numerical experiments using real data as an example. We plan to continue our experimental studies by considering the domain-specific issues of IR and developing guidelines for the use of predictive models in IR.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP19K03005 and JP22K02832.

References

[1] C. Brooks and C. Thompson, "Predictive Modelling in Teaching and Learning," Handbook of Learning Analytics, pp. 61–68, SoLAR, 2017.

[2] A. Parnell, D. Jones, A. Wesaw, and D. C. Brooks, "Institutions' Use of Data and Analytics for Student Success," NASPA, AIR and EDUCAUSE, 2018.

[3] K. Deb, Multi-Objective Optimization using Evolutionary Algorithms, New York, USA: John Wiley & Sons, 2001.

[4] N. Kondo, T. Matsuda, Y. Hayashi, H. Matsukawa, M. Tsubakimoto, Y. Watanabe, S. Tateishi, and H. Yamashita, "An Approach for Academic Success Predictive Modeling based on Multi-objective Genetic Algorithm", International Journal of Institutional Research and Management, Vol.5, No.1, pp.31–49, 2021.