Sharpe Ratio Based Index for Building Fault Prediction Model

Masateru Tsunoda, Akito Monden, and Ken-ichi Matsumoto
Graduate School of Information Science
Nara Institute of Science and Technology
Nara, Japan
{masate-t, akito-m, matumoto}@is.naist.jp

***Abstract*—The fault-prone modules prediction model which is built on the fit dataset often does not suit the test data well. As a result, prediction accuracy of the model on the test dataset is lower than the fit dataset. To prevent depression of the accuracy on the test dataset, we propose Sharpe ratio based index. It identifies stable and strong relationships between a response and explanatory variable to prevent that.**

Keywords-Overfitting; Sharpe ratio; variable selection; area under the curve; coefficient of variation

I. INTRODUCTION

Fault-prone modules prediction model is useful to find software faults effectively. The model is built based on a dataset collected in past projects (fit dataset), and fault-prone modules are predicted by applying the model to a dataset collected from ongoing project (test dataset). When building the model, variable selection and setting proper parameters of the model are performed to enhance prediction accuracy. Some indices such as correlation coefficient and AUC (area under the curve) [2] are used to perform that.

However, the model which is built on the fit dataset often does not suit the test data well. As a result, prediction accuracy on the test dataset is lower than the fit dataset. This is because there are differences of characteristics between the fit and test dataset. We assume that a dataset consists of a stable and uncertain (noisy) part. The characteristic of the stable part seldom varies, but that of the uncertain part often does. So, the characteristic of the stable part of the fit dataset is similar to the test dataset, but that of the uncertain part of fit dataset is greatly different from the test dataset.

If the stable part and the uncertain part are identified, and the prediction model is built using the stable part only, the depression of the accuracy on the test dataset is expected to be prevented. To identify the stable and uncertain part, we focus on a variance of the index used to build the model (e.g. a variance of correlation coefficient). The variance is computed by the resampling method [1]. The method regards a sample as a population, and extracts cases from the sample many times to estimate the distribution of the population. The uncertain part is identified by the variance. For example, if a variance of a correlation coefficient between a response and explanatory variable is large, the relationship is regarded as uncertain.

In addition, it is not ignorable the value of the index. For instance, if the variance of the correlation coefficient is small, but the average of the coefficient also is small, it should not be used as the explanatory variable. To consider both the value and variance of the index, we propose Sharpe ratio based index. The Sharpe ratio is originally used to evaluate performance of a portfolio (combined financial products). It takes into account not only profit but also risk. When profit is high but risk (standard deviation) is also high, the value is low. Similarly, the Sharpe ratio based index considers both the value and variance of the index. For instance, a candidate explanatory variable whose correlation coefficient for a response variable is large, and the variance of that calculated by the resampling method is small, the variable is selected as the explanatory variable.

II. MODEL BUILDING PROCEDURE

Using the proposed index, a fault-prone module prediction model is built with the following procedure.

Step 1: Cases are extracted from the fit dataset by bootstrapping. It is one of the resampling methods, and it randomly extracts cases from the fit dataset. The number of extracted cases is the same as the fit dataset, and a case is extracted more than once (sampling with replacement). For example, when the fit dataset includes cases $\{M_1, M_2, M_3, M_4\}$, bootstrapping extracts $\{M_1, M_1, M_2, M_3\}$ or $\{M_1, M_2, M_2, M_4\}$. Generally, the extraction is repeated 1000 times.

Step 2: Average and variance of an index used to build a prediction model are computed, using datasets extracted in step 1. For instance, a correlation coefficient between a response and explanatory variable is calculated on each extract dataset, and as a result, 1000 values of them are made. Using the values, average and variance of the correlation coefficient are computed.

Step 3: A fault-prone module prediction model is built using the Sharpe ratio based index. The Sharpe ratio based index c is calculated by $c = a / d$. In the equation, a denotes average of the index, and d denotes standard deviation of it. When average is large, standard deviation is also large. To normalize standard deviation, coefficient of variation (CV) is applicable, instead of the standard deviation. CV is standard deviation divided by average (That is, c is computed by a^2/d).

III. EXPERIMENT

A. *Identifying stable and strong relationship*

We confirmed whether proposed indices identify variables which have stable and strong relationships to a

TABLE I. *AE*, *RE*, AND *IRE* OF CORRELATION, *SI*, AND *SICV*.

	AE	*RE*	*IRE*
Correlation	4.3	0.51	0.69
SI	4.16	0.36	0.52
SICV	3.78	0.34	0.51

response variable or not in the experiment. We used MW1 project dataset which is offered by NASA Metrics Data Program (MDP) [3]. The dataset includes 403 cases and 37 metrics. The rate of fault-prone modules is 7.7%. Experimental procedure is as follows.

1. The dataset is randomly divided into a fit dataset and a test dataset.
2. Using the fit dataset, average and variance of Spearman's rank correlation coefficient between a response and candidate explanatory variable are calculated by bootstrapping.
3. Sharpe ratio based index (*SI*) and Sharpe ratio based index using CV (*SICV*) are computed.
4. Using the test dataset, correlation coefficient between a response and candidate explanatory variable is calculated.

According to magnitude, we ranked correlation coefficient, *SI*, and *SICV* on the fit dataset. Similarly, correlation coefficient on the test dataset was ranked. Next, we calculated absolute error (*AE*), relative error (*RE*), and inverse relative error (*IRE*) of the rank by $AE = |f - t|$, $RE = |f - t| / t$, $IRE = |f - t| / f$. In the equations, f denotes the rank on the fit dataset, and t denotes the rank on the test dataset. When *AE*, *RE*, and *IRE* are lower, the relationships are more stable and stronger. We used ranking to evaluate the performance of the indices, because we assume forward or backward variable selection is performed with the index based ranking to choose explanatory variables.

Table I shows *AE*, *RE*, and *IRE* of correlation coefficient, *SI,* and *SICV*. *SICV* has the lowest values, and we conclude that *SICV* is most effective to identify stable and strong relationships. The reason that correlation coefficient did not work well is that it identifies strong relationships but did not do stable ones between a response and explanatory variable.

B. *Variable selection for prediction*

Next, we evaluated the accuracy of the fault-prone modules prediction when variable selection based on correlation coefficient, Sharpe ratio based index (*SI*), and Sharpe ratio based index using CV (*SICV*) are applied. After the variable selection, fault-prone modules are predicted by analogy-based method [4]. To evaluate the accuracy of the prediction, we used AUC (area under the curve) [2]. AUC is area under the ROC (receiver operating characteristic) curve, and the range of value is [0, 1]. High value of AUC indicates the accuracy of the model is high. We used MW1 project dataset. Experimental procedure is as follows. In the procedure, step 1 to 3 is same as the procedure of the preceding experiment.

4. A candidate explanatory variable whose value of the criterion is the smallest is removed (backward elimination) from the fit and test dataset.

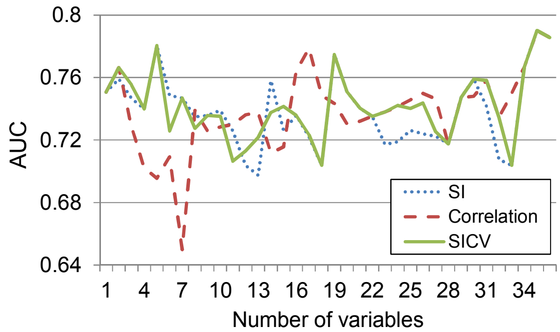

Figure 1. Relationships between accuracy and the number of variables.

5. Fault-prone modules in the test dataset are predicted, and ROC is calculated based on the result.
6. Until candidate explanatory variables exists in the dataset, Step 4 and 5 are repeated. After that, we go back to step 2 to apply other indices.

Fig. 1 shows AUC of each criterion. When almost all candidate explanatory variables are used, prediction accuracy was highest, and there is no difference between the criteria. So, both *SI* and *SICV* did not show explicit effect in variable selection. However, when the number of explanatory variables was small and correlation coefficient was used as the criterion, prediction accuracy was lower than others. The result suggests correlation coefficient may not work well when the number of explanatory variables was small.

IV. CONCLUSIONS

We propose Sharpe ratio based index (*SI*) and Sharpe ratio based index using coefficient of variation (*SICV*) for fault-prone module prediction. Our experiment showed *SICV* is effective to identify stable and strong relationships between a response and explanatory variable. But they did not improve accuracy of fault-prone module prediction in the experiment. Our future work is to confirm *SI* and *SICV* when using other datasets and prediction methods.

ACKNOWLEDGMENT

This work is being conducted as a part of the StagE project, The Development of Next-Generation IT Infrastructure, and Grant-in-aid for Young Scientists (B), 22700034, 2010, supported by the Ministry of Education, Culture, Sports, Science and Technology. We would like to thank Dr. Naoki Ohsugi for offering the collaborative filtering tool.

REFERENCES

[1] P. Good: Resampling Methods: A Practical Guide to Data Analysis, Birkhäuser Boston, 2005.

[2] J. Hanley, and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, no.143, pp.29-36, 1982.

[3] The National Aeronautics and Space Administration: Metrics Data Program, http://mdp.ivv.nasa.gov.

[4] M. Shepperd, and C. Schofield, "Estimating software project effort using analogies," IEEE Transactions on Software Engineering, vol.23, no.12, pp.736-743, 1997.