Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 157 (2019) 229–237

Procedia Computer Science

www.elsevier.com/locate/procedia

4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), 12-13 September 2019

Modified Average of the Base-Level Models in the Hill-Climbing Bagged Ensemble Selection Algorithm for Credit Scoring

Tri Handhika[a,*], Achmad Fahrurozi[a], Revaldo Ilfestra Metzi Zen[b], Dewi Putrie Lestari[a], Ilmiyati Sari[a], Murni[a]

[a] *Computational Mathematics Study Center, Gunadarma University, Margonda Raya St. No. 100, Depok 16424, Indonesia*
[b] *Big Data Division, Metra Digital Media, Wisma Aldiron Dirgantara 2nd Floor Suite 202-209, South Jakarta 12780, Indonesia*

Abstract

Performance of credit scoring model is a main concern for financial institutions in determining the credit risk of credit applicants. Credit score will be one of basis for the lender to make a decision, approved or rejected, for any credit applications. There are many methods and approaches that have been modeled for this problem. This study tries to explore further the Hill-Climbing Bagged Ensemble Selection (HCES-Bag) algorithm which has the best performance for credit scoring model as has been analyzed comprehensively in the research conducted by Lessmann et al.[1]. We modify some average formulas for the base-level models to find out the opportunity for improving the performance of credit scoring model as measured by several performance indicators. Experiment with German Credit Data from the UCI Machine Learning Repository by first using Multivariate Adaptive Regression Splines (MARS) model for features selection demonstrates that the modification average does not affect credit scoring model performance significantly. However, some of them make the credit scoring model become more efficient because we can obtained same level of credit scoring model performances by using only smaller number of base-level models.

Keywords: Heterogeneous ensemble classifier ; hill-climbing ; ensemble selection ; bagging ; average ; credit scoring.

1. Introduction

For the last few years, credit industry has developed rapidly[2]. It is indicated by the rise of financial technology businesses that make people easy to apply for credit. However, this facility needs to be watched out by the lenders by first determining the credit risk of credit applicants known as credit scoring. Credit scoring as one of popular research topics in the field of financial risk management can be used as a credit risk evaluation method to prevent credit default by generating a score that describes creditworthiness of credit applicants[3]. Credit scoring helps financial institutions as lenders to measure not only credit applicants' ability to pay, but also their willingness to pay. Based on

[*] Corresponding author. Tel.: +62-811-899-1399.
E-mail address: trihandika@staff.gunadarma.ac.id

1877-0509 © 2019 The Authors. Published by Elsevier B.V.

10.1016/j.procs.2019.08.162

that characteristics, credit scoring model is built to classify credit applicants as good or bad borrowers. This prediction will be a basis for lenders to approve or reject the applications. The main idea of credit scoring model is to identify the effect of features, either ability or willingness to pay of credit applicants, that affect their payment behavior as well as their credit default risk, besides other demographic features.

Credit scoring model is classified as supervised learning problem in the context of machine learning. There are two fundamental stages to build appropriate credit scoring model, i.e. features selection and classification. There are also some of techniques to model credit risk that have been developed, such as expertise scoring model, statistical model, artificial intelligence, or hybrid that is combination of some techniques for each stage or even mixed on only one stage, i.e. classification, known as ensemble classifier[4]. In this paper, we use hybrid technique where statistical model applied for features selection while ensemble classifier for classification. However, the discussion is limited to the classification stage only. The statistical model used in the features selection stage is Multivariate Adaptive Regression Splines (MARS)[5].

Ensemble classifiers were selected at the classification stage because we had previously analyzed the results of research conducted by Lessmann et al.[1] They found that the Hill-Climbing Bagged Ensemble Selection (HCES-Bag) algorithm, one of static direct heterogeneous ensemble classifiers, is the best classifier for credit scoring model from a spectacular experiment including combination of 8 data sets with various features and number of cases, 41 algorithms from all of types of classifiers with thousands of base-models, and 6 different performance indicators, and of course, some of number-fold cross validation. The heterogeneous ensemble classifier is the best type of classifiers for all of performance indicators. Although HCES-Bag is the best on only 2 performance indicators, but it was ranked first on the grand average of all performance indicators[1].

The idea of heterogeneous ensemble classifiers is that different algorithms can complement each other because they have different perspectives on the same data[1]. These algorithms are combined through a weighted average or voting of base-level models[6]. It means that the objective performances of heterogeneous ensemble classifiers, including HCES-Bag, depend on the right weighted formula. The original HCES (without bootstrap aggregating (bagging)) algorithm starts with an ensemble of the best base models and growing continues by adding the base models that increase ensemble performance using forward stepwise selection from the base model library in a fully-enumerative procedure until it stops improving as explained comprehensively by Lessmann et al.[1] in their online appendix. To reduce overfitting, bagging is used for improving the ensemble selection by running the HCES algorithm multiple times to the only different bootstrap samples from the library of base models[6]. In this paper, we tried to modify the average of the base-level models in the HCES-Bag algorithm to find out the opportunity for improving the performance of credit scoring model.

2. Literature Review

2.1. Features Selection with Multivariate Adaptive Regression Splines

Some of statistical and artificial intelligence techniques have been widely applied to develop the features selection stage in constructing a credit scoring model, such as Logistic Regression[7], Genetic Algorithm[8], Neighborhood Rough Set[9], and Multivariate Adaptive Regression Splines (MARS)[10]. In this paper, we used MARS model for features selection because of its power ability to generalize the results of high-dimensional data processing[4]. The MARS model fits the relationship between a set of independent variables (features) and a dependent variable by dividing the space of all features into multiple values as known as knots in order to fit a spline function between these knots[11]. MARS model searches over all possible univariate locations in across interactions among all features by using the basis function which is analog to the splines[12]. During the searching which is performed in an iterative process, an increasingly larger number of basis functions are added to the model to minimize a lack-of-fit criterion such that the most important features are determined simultaneously. The features as well as the knot locations having the most contribution to the model are selected first. At the end of each iteration, the indication of an interaction is examined for any possible model improvements. The general MARS model equation is expressed as[12]:

$$f(x) = \beta_0 + \sum_{r=1}^{R} \beta_r h_r(x), \tag{1}$$

where $\beta_0, \beta_1, \beta_2, \cdots, \beta_R$ are the coefficients of the model. They are estimated to yield the best fit to the data with R sub-regions or basis functions in the model as defined as spline basis function $h_r(x)$[12].

2.2. Classification with Heterogeneous Ensemble Classifier: Hill-Climbing Bagged Ensemble Selection

Lessmann et al. have been group different classifiers into three families, i.e. individual, homogeneous ensemble, and heterogeneous ensemble classifiers[1]. Artificial Neural Network (ANN)[10], Classification and Regression Tree (CART)[13], and Support Vector Machines (SVM)[9] are examples of individual classifier algorithms which are used in constructing a credit scoring model. In addition, each of Random Forest[14] and Hill-Climbing Bagged Ensemble Selection (HCES-Bag)[1] is an example of homogeneous and heterogeneous ensemble classifiers, respectively, applied in credit scoring. Heterogeneous ensemble classifier is more appropriate than a single complex classifier in credit scoring model because of the data derived from different financial institutions have their own properties[15]. It combines multiple classification models from different classification algorithms that incorporates all individual classifiers and homogeneous ensemble classifiers. The combination of different classifier algorithms can increase the performance of classification as long as the diversity does not reduce the performance of each classifier[16]. Heterogeneous ensemble classifier involves three stages, creating a set of base models, ensemble selection, and combining their predictions using some pooling mechanism[1]. There are two strategies in ensemble selection depending on the selection step management, static or dynamic. If the base model searching only performed once then ensemble selection is called as static, otherwise is called as dynamic. If the base model chosen based on the increasing performance of the heterogeneous ensemble classifier then it is called as a direct approach. Conversely, indirect approaches focused on the other determinant of ensemble success such as diversity among base models[1].

Let a composite heterogeneous ensemble prediction of T base-level models for i-th credit applicant who has features $\mathbf{x}_i$ is defined as follow:

$$E(\mathbf{x}_i, \mathbf{M}) = \sum_{t=1}^{T} w_t M_t(\mathbf{x}_i), \tag{2}$$

where $M_t(\mathbf{x}_i)$ predicts the individual credit score of each base model from different classification algorithms in $\mathbf{M} = (M_1, M_2, \cdots, M_T)$ with their own weight w_t. Each model has different predictive output scale so it needs to be calibrated using an appropriate method, i.e. the Platt's method[17]. In this paper, we tried to modify the average formula on equation (2) as well as to determine the appropriate formula for each weight w_t. This experiment is inspired by Lessmann et al. who compute the weights $w_t, t = 1, 2, \cdots, T$ by using the predictive accuracy of each base model in terms of some performance indicators[18]. We will know what the effect of this modification to the credit scoring model performance. There are three types of performance indicators that considered in this study, i.e. the Percentage Correctly Classified (PCC) to assess the correctness of the categorical credit score predictions, the Area Under the Curve (AUC) of Receiver Operating Characteristic (ROC) to assess the discriminatory ability of the categorical credit score, and the Brier Score (BS) to assess the accuracy of the credit score's probability predictions[1].

HCES-Bag algorithm is one of static direct heterogeneous ensemble classifiers. It starts with Top-T approach that is an ensemble of the best T base models and then growing continues by adding the base model searched once from the library that increases ensemble performance until it stops improving. Multiple inclusion of the same base model is allowed for this algorithm[1]. It forms all possible candidate ensembles of $T + 1$ members in a fully-enumerative procedure, and examines which ensemble that gives higher performance, the original ensemble of size T or the augmented ensemble of size $T + 1$. The augmented ensemble which increases performance the most covers the original ensemble. The simple forward stepwise selection procedure is usually used in the original HCES algorithm to develop $\mathbf{M}$ on equation (2), but sometimes causes overfitting so that reducing the credit scoring model performance. To reduce overfitting, bagging is then applied by running HCES algorithm multiple times to the only different bootstrap samples from the library of base models for improving the ensemble selection[6]. The HCES-Bag prediction is then computed by average formula on equation (2).

3. Research Methodology

In this study, we propose some modifications of average formulas for the base-level models in the HCES-Bag algorithm applied for credit scoring. We first build a library of models from both individual and homogeneous ensemble classification algorithms, i.e. Stochastic Gradient Descent (SGD)[19], Gradient Boosting Classifier (GBC)[14], Decision Trees (DT)[20], Random Forests (RF)[21], Extremely Randomized Trees (ERT)[22], pipelining K-Means[23] and Logistic Regression[7] (PKMLR), and pipelining Nyström[24] and Logistic Regression[7] (PNLR). To form a library consisting of many models, the several parameter values of each classifier are changed. For example, we can determine one parameter of RF, the criteria for grouping "gini" or "entropy", so that there are two different models in the library. From these two models, we can also develop several more models by determining other different parameters.

The original credit scoring prediction is obtained by simple averaging the credit score of each selected model uniformly in the HCES-Bag classifier with T_0 base-level models on equation (2) denoted by $E_0\left(\mathbf{x}_i, \mathbf{M}^{(0)}\right)$. In this paper, we tried to modify the average formulas into five new formulas denoted by $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right), E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right), E_3\left(\mathbf{x}_i, \mathbf{M}^{(3)}\right), E_4\left(\mathbf{x}_i, \mathbf{M}^{(4)}\right)$, and $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$ for different T_1, T_2, T_3, T_4 and T_5 base-level models, respectively. Modification consists of determining weights for each model or re-formulating the average on equation (2). $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$ is formulated based on the basic principle of building an ensemble classifier where the more base models drawn, the higher degree of model compatibility with the credit data. Therefore, a model that was sampled most in the HCES-Bag algorithm has the greatest weight in the $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$. The general formula of $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$ can be expressed as on equation (3):

$$E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right) = \frac{\sum_{t=1}^{T_1}\left(w_t^2 M_t^{(1)}\right)}{\sum_{t=1}^{T_1} w_t^2} \tag{3}$$

where w_t is the number of each model $M_t^{(1)}, t = 1, 2, \cdots, T_1$ in the HCES-Bag algorithm.

We formulated the average $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right), E_3\left(\mathbf{x}_i, \mathbf{M}^{(3)}\right)$, and $E_4\left(\mathbf{x}_i, \mathbf{M}^{(4)}\right)$ following Lessmann et al.[18] as mentioned in Subsection 2.2 in terms of PCC, AUC, and BS, respectively, by substituting the validation set with k-fold cross validation procedure on the training set. The last modified average we propose is the geometric average, $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$, by considering that the individual prediction of each base-level model is converted into interval [0, 1]. The geometric average of the base-level models is expressed as on equation (4):

$$E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right) = \prod_{t=1}^{T_5}\left(1 + M_t^{(5)}\right)^{\frac{1}{T_5}} - 1. \tag{4}$$

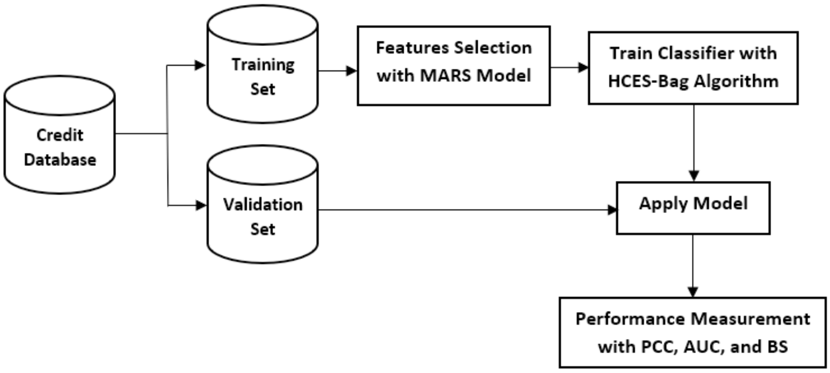

Credit Database
Training Set
Validation Set
Features Selection with MARS Model
Train Classifier with HCES-Bag Algorithm
Apply Model
Performance Measurement with PCC, AUC, and BS

Fig. 1: General design of experiment in this research.

Fig. 1 illustrates the design of experiment for this research where credit database is divided into two groups, i.e. training and validation sets. Credit data in the training set is then processed to select significant features. In this study, we used MARS model on equation (1) as a feature selector. Furthermore, the HCES-Bag algorithm as a static direct heterogeneous ensemble classifier is applied such that can be used to classify the good or bad credit applicants in the validation set. Performance of the model generated then measured by some performance indicators, i.e. PCC, AUC, and BS, as mentioned in Subsection 2.2. All of the average formulas introduced in this paper, $E_n\left(\mathbf{x}_i, \mathbf{M}^{(n)}\right)$ for $n = 0, 1, 2, 3, 4, 5$, are applied in the HCES-Bag algorithm for credit scoring as shown on Fig. 2. The HCES-

Bag algorithm starts with determining the fraction of pruning some base models (p)[25], the fraction of samples size bootstrapped (b), the number of ensembles (N), the maximum number of models in each ensemble (Max), and the initial number of models with the best performances in each ensemble (T). Furthermore, prune the models with a fraction p that had previously been run with K-fold cross-validation based on the credit data in the training set which have bad performances (PCC, AUC, or BS). Bootstrap samples are then performed N times from the selected models after pruning where its size depending on a fraction b. The process is continued by selecting T base models with the best performances (PCC, AUC, or BS) for the initial ensemble. Finally, ensemble selection process as explained in Subsection 2.2 is run with all of modified average formulas for each performance indicator as long as the number of models in each ensemble does not exceed the Max.

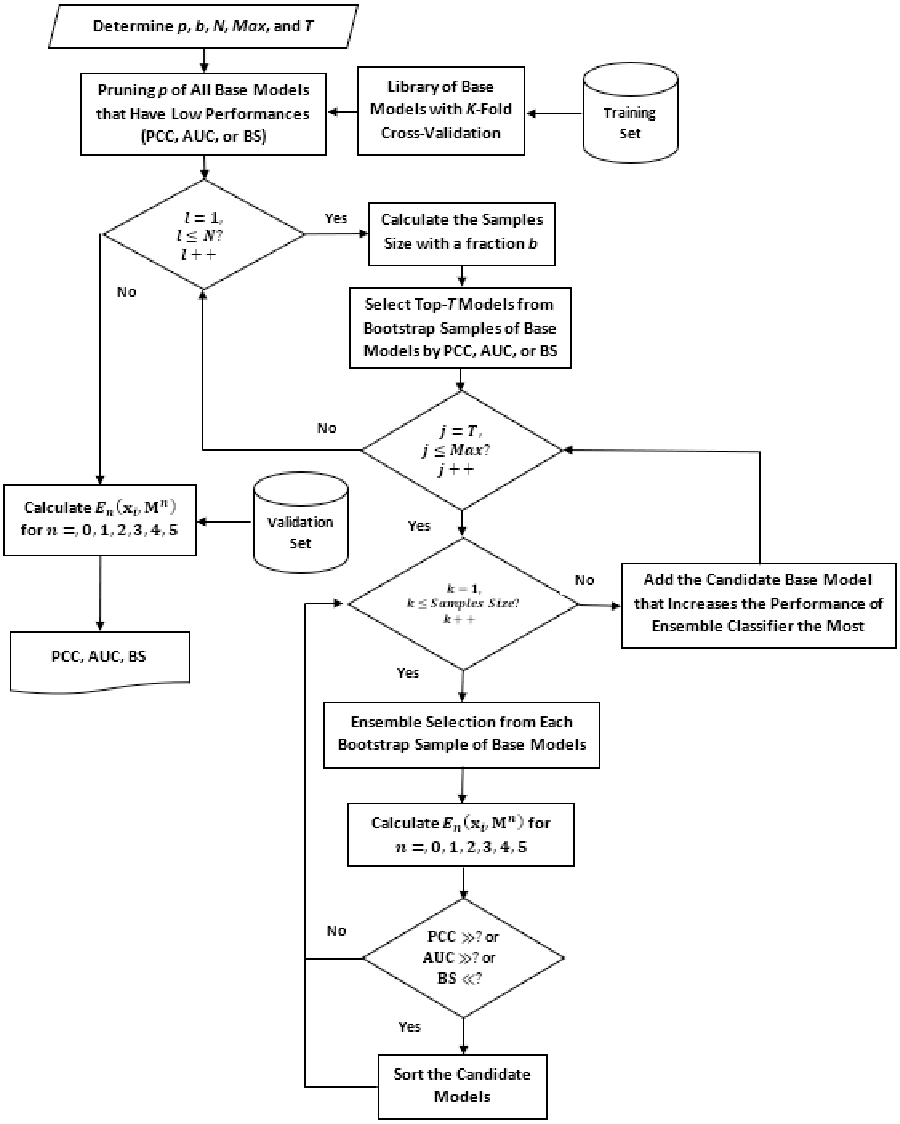

Determine p, b, N, Max, and T
Pruning p of All Base Models that Have Low Performances (PCC, AUC, or BS)
Library of Base Models with K-Fold Cross-Validation
Training Set
$l = 1$, $l \leq N?$ $l++$
Yes
Calculate the Samples Size with a fraction b
No
Select Top-T Models from Bootstrap Samples of Base Models by PCC, AUC, or BS
No
$j = T$, $j \leq Max?$ $j++$
Calculate $E_n(\mathbf{x}_i, \mathbf{M}^n)$ for $n =, 0, 1, 2, 3, 4, 5$
Validation Set
Yes
$k = 1$, $k \leq Samples\ Size?$ $k++$
No
Add the Candidate Base Model that Increases the Performance of Ensemble Classifier the Most
PCC, AUC, BS
Yes
Ensemble Selection from Each Bootstrap Sample of Base Models
Calculate $E_n(\mathbf{x}_i, \mathbf{M}^n)$ for $n =, 0, 1, 2, 3, 4, 5$
No
PCC ≫? or AUC ≫? or BS ≪?
Yes
Sort the Candidate Models

Fig. 2: Design of experiment for the HCES-Bag algorithm with modified average of the base-level models in this research.

4. Results and Discussions

In this Section, we conduct an experiment of modified average formulas in the HCES-Bag algorithm for credit scoring using German Credit Data from the UCI Machine Learning Repository. This data is collected from a sample of credit applicants of size 1000 which consist of 700 creditworthy applicants and 300 applicants who should be rejected. There are 20 features consisting of 7 numerical and 13 categorical attributes. The numerical attributes are duration in month, credit amount, installment rate in percentage of disposable income, present residence since, age in years, number of existing credits at this bank, and number of people being liable to provide maintenance for. While, the categorical attributes are status of existing checking account, credit history, purpose, savings account/bonds, present employment since, personal status and sex, other debtors/guarantors, property, other installment plans, housing, job, telephone, and foreign worker. We follow the design of experiment on Fig. 1 by splitting this data proportionally using stratified random sampling into two groups, i.e. training and validation sets of size 800 and 200, respectively. There are only three features that were selected by MARS model to distinguish the good or bad credit applicants on the classification stage, i.e. status of existing checking account, duration in month, and credit amount.

In this experiment, we have a total of 2556 base models in our library with $p = 75\%, b = 25\%, N = 20, Max = 25, T = 5$, and $K = 3$, see Fig. 2. The left side of Table 1 (Performance Indicator column) presents performance of each $E_n\left(\mathbf{x}_i, \mathbf{M}^{(n)}\right)$ for $n = 0, 1, 2, 3, 4, 5$ in the HCES-Bag algorithm, where the best measurement for each performance indicator and each $E_n\left(\mathbf{x}_i, \mathbf{M}^{(n)}\right)$ is in the grey area. Note that bold number is used for the best measurement for each performance indicator. The HCES-Bag algorithm with Top-T approached by AUC for both original average, $E_0\left(\mathbf{x}_i, \mathbf{M}^{(0)}\right)$, and modified average, $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$, generate the highest PCC of all experiments conducted, i.e. 73%. It means that the model succeeded to determine exactly 146 of the 200 applicants in the validation set who classified as either good or bad credit applicants. We can use this model to determine the credit score of new credit applicants with an accuracy of 73%. The highest AUC is obtained by the HCES-Bag algorithm with Top-T approached by BS for modified average, $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$, i.e. 75.869%. The lowest BS is obatined by the HCES-Bag algorithm with Top-T approached, again, by BS for modified average, $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right)$, which computing the weights of equation (2) by using PCC, i.e. 17.629%. Based on the results, we know that the modification average does not affect credit scoring model performance significantly. Moreover, Top-T approached by BS has best performances (PCC, AUC, and BS) for almost all modified average in the HCES-Bag algorithm while Top-T approached by PCC is not recommended.

Table 1: Performance of each $E_n\left(\mathbf{x}_i, \mathbf{M}^{(n)}\right)$ and the number of base models used in the HCES-Bag algorithm.

n	**Top-*T* Approach**	**Performance Indicator**			**Individual/Homogeneous Ensemble Classifier Algorithm**							**Total**
		PCC	**AUC**	**BS**	**SGD**	**GBC**	**DT**	**RF**	**ERT**	**PKMLR**	**PNLR**	
0	PCC	0.71500	0.75440	0.17722	0	115	19	60	165	2	0	361
	AUC	**0.73000**	0.75381	0.17872	0	313	0	116	71	0	0	500
	BS	0.72000	0.74583	0.17732	0	270	56	51	123	0	0	500
1	PCC	0.70000	0.75024	0.1790	0	76	11	50	143	2	0	282
	AUC	**0.73000**	0.75762	0.17791	0	117	0	73	38	0	0	228
	BS	0.72500	0.75238	0.17763	0	39	4	42	35	0	0	**120**
2	PCC	0.71000	0.75369	0.17689	0	94	10	62	188	2	0	356
	AUC	0.71500	0.75595	0.17711	0	78	0	51	187	0	0	316
	BS	0.72500	0.75786	**0.17629**	0	70	9	52	209	0	0	340
3	PCC	0.71000	0.75321	0.17727	0	338	10	31	115	2	0	496
	AUC	0.72500	0.75750	0.17814	0	280	0	151	49	0	0	480
	BS	0.72500	0.75821	0.17655	0	200	181	42	61	0	0	484
4	PCC	0.72000	0.75048	0.17726	0	123	10	31	49	2	0	215
	AUC	0.72500	0.75369	0.1777	0	126	0	42	48	0	0	216
	BS	0.72000	0.75512	0.17671	0	117	18	44	46	0	0	225
5	PCC	0.71500	0.75405	0.21008	0	116	14	44	164	2	0	340
	AUC	0.72500	0.75774	0.24878	0	287	0	133	70	0	0	490
	BS	0.72500	**0.75869**	0.18224	0	48	5	42	25	0	0	**120**

We explore the results further related to the base models selected in each experiment of the HCES-Bag algorithm as shown on the right side of Table 1 (Individual/Homogeneous Ensemble Classifier Algorithm column). The base models based on the SGD and the PNLR algorithms are never selected in the ensemble selection for all experiments. The base models based on the DT algorithm are never selected in the ensemble selection for experiments which Top-T

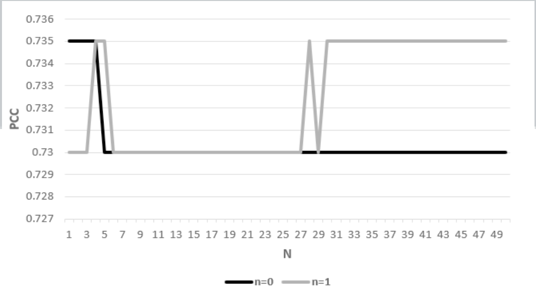
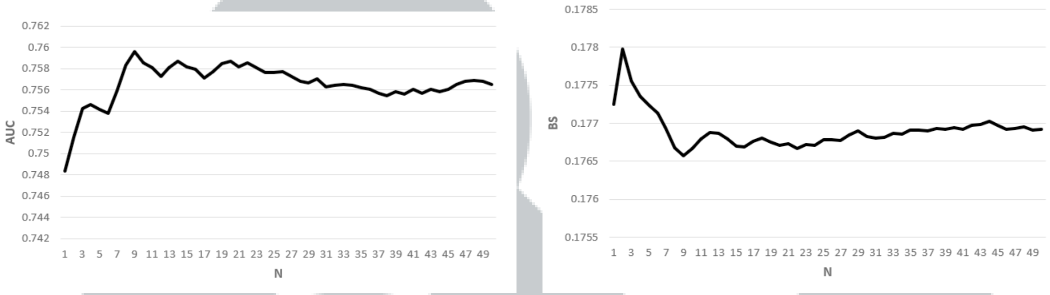


Fig. 3: The PCC of the HCES-Bag algorithm for several number of ensembles using $E_0\left(\mathbf{x}_i, \mathbf{M}^{(0)}\right)$ and $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$ with Top-T approached by AUC.



(a) The AUC of the HCES-Bag algorithm using $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$.

(b) The BS of the HCES-Bag algorithm using $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right)$.

Fig. 4: The AUC and BS of the HCES-Bag algorithm for several number of ensembles using $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$ and $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right)$, respectively, with Top-T approached by BS.

approached by AUC. The base models based on the PKMLR algorithm are never selected in the ensemble selection for experiments which Top-T approached by AUC as well as BS. The GBC algorithm has the largest number of base models used in this experiment, except for $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right)$ and Top-T approached by PCC for both $E_0\left(\mathbf{x}_i, \mathbf{M}^{(0)}\right)$ and $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$. This is inline with the previous study which stated that GBC performs very well for imbalanced data sets like German Credit Data[14]. Top-T approached by BS, especially for $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$ and $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$, used the smallest number of base models in the experiment, i.e. 120 base models. Based on the results of Table 1, we recommend the HCES-Bag algorithm with Top-T approached by BS for modified average, $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$, to be a credit scoring model for German Credit Data because of its good performance by using only smaller number of base-level models.

The experiment was continued by changing the number of ensembles used in each experiment which have the best performance from Table 1. Fig. 3 shows that the more number of ensembles, the modified average, $E_1\left(\mathbf{x}_i, \mathbf{M}^{(1)}\right)$, has a greater probability to increase the PCC than the original average, $E_0\left(\mathbf{x}_i, \mathbf{M}^{(0)}\right)$ with Top-T approached by AUC. Moreover, the more number of ensembles, the AUC of the HCES-Bag algorithm using $E_5\left(\mathbf{x}_i, \mathbf{M}^{(5)}\right)$ with Top-T approached by BS will be consistent around 75.5% as shown on Fig. 4(a). The BS of HCES-Bag algorithm using $E_2\left(\mathbf{x}_i, \mathbf{M}^{(2)}\right)$ with Top-T approached by BS will be around 17.7% as the increasing number of ensembles, see Fig. 4(b).

5. Conclusions and Future Works

In this paper, we have been formulated five modified average formulas for the base-level models in the HCES-Bag algorithm for credit scoring. They are implemented to the German Credit Data by first using MARS model to select the significant features to distinguish the good or bad credit applicants on the classification stage. Only 3 of 20 features provided by German Credit Data that were selected by MARS model, i.e. status of existing checking account, duration in month, and credit amount. The experiments are conducted by using this three significant features in terms of three performance indicators, i.e. PCC, AUC, and BS. The results show that the credit scoring model performance does not affected significantly by this modification, but it makes the credit scoring model become more efficient because we can obtained same level of credit scoring model performances by using only smaller number of base-level models. To generalize this results, the same experiment can be conducted to the other credit data sets. Furthermore, it can be developed by using the other base models such as, Support Vector Machine (SVM), Classification and Regression Trees (CART), Artificial Neural Network (ANN), and so on. The other performance indicators like H-measure or Partial Gini index can also be used for future works.

Acknowledgements

This research was supported by PUPT 2019 Research Grant, Ministry of Research, Technology, and Higher Education, Indonesia to the Gunadarma University No. 009.13/LP/UG/V/2019.

References

1. Lessmann, S., Baesens, B., Seow, H.V., Thomas, L.C.. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research* 2015;**247**:124–136.
2. Chuang, C.L., Lin, R.H.. Constructing a reassigning credit scoring model. *Expert Systems with Applications* 2009;**36**:1685–1694.
3. Louzada, F., Ara, A., Fernandes, G.B.. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science* 2016;**21**:117–134.
4. Li, X.L., Zhong, Y.. An overview of personal credit scoring: Techniques and future work. *International Journal of Intelligence Science* 2012; **2**:181–189.
5. Friedman, J.H.. Multivariate adaptive regression splines. *The Annals of Statistics* 1991;**19**(1):1–141.
6. Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A.. Ensemble selection from libraries of models. *In Proceedings of the* 21^{th} *International Conference on Machine Learning* 2004;:18–25.
7. Dong, G., Lai, K.K., Yen, J.. Credit scorecard based on logistic regression with random coefficients. *Procedia Computer Science* 2012; **1**:2463–2468.
8. Šušteršič, M., Mramor, D., Zupan, J.. Consumer credit scoring models with limited data. *Expert Systems with Applications* 2009;**36**(3):4736–4744.
9. Yao, P.. Hybrid classifier using neighborhood rough set and svm for credit scoring. *In Proceedings of 2009 International Conference on Business Intelligence and Financial Engineering* 2009;:138–142.
10. Lee, T.S., Chen, I.F.. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications* 2005;**28**(4):743–752.
11. Zhang, W., Goh, A.T.C.. Multivariate adaptive regression splines and neural network models for prediction of pile drivability. *Geoscience Frontiers* 2016;**7**:45–52.
12. Oduro, S.D., Metia, S., Duc, H., Hong, G., Ha, Q.P.. Multivariate adaptive regression splines models for vehicular emission prediction. *Visualization in Engineering* 2015;**3**(13):1–12.
13. Lee, T.S., Chiu, C.C., Chou, Y.C., Lu, C.J.. Mining the customer credit using classification and regression tree and multivariate adaptive regression splines. *Computational Statistics & Data Analysis* 2006;**50**(4):1113–1130.
14. Brown, I., Mues, C.. An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications* 2012;**39**(3):3446–3453.
15. Xia, Y., Liu, C., Li, Y., Liu, N.. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications* 2017;**78**:225–241.
16. Abellán, J., Castellano, J.G.. A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications* 2017;**73**:1–10.
17. Platt, J.C.. Probabilities for sv machines. *In Advances in Large Margin Classifiers* 2000;:61–74.
18. Lessmann, S., Sung, M.C., Johnson, J.E.V., Ma, T.. A new methodology for generating and combining statistical forecasting models to enhance competitive event prediction. *European Journal of Operational Research* 2012;**218**:163–174.
19. Tomczak, J.M.a.. Classification restricted boltzmann machine for comprehensible credit scoring model. *Expert Systems with Applications* 2015;**42**:1789–1796.
20. Sohn, S.Y., Kim, J.W.. Decision tree-based technology credit scoring for start-up firms: Korean case. *Expert Systems with Applications* 2012; **39**:4007–4012.
21. Breiman, L.. Random forests. *Machine Learning* 2001;**45**:5–32.
22. Geurts, P., Ernst, D., Wehenkel, L.. Extremely randomized trees. *Machine Learning* 2006;**63**:3–42.
23. Xiao, H., Xiao, Z., Wang, Y.. Ensemble classification based on supervised clustering for credit scoring. *Applied Soft Computing* 2016; **43**:73–86.

24. Yang, T., Li, Y.F., Mahdavi, M., Jin, R., Zhou, Z.H.. Nyström method vs random fourier features: A theoretical and empirical comparison. *In Proceedings of the 25th International Conference on Neural Information Processing Systems* 2012;**1**:476–484.

25. Partalas, I., Tsoumakas, G., Vlahavas, I.. An ensemble uncertainty aware measure for directed hill climbing ensemble pruning. *Machine Learning* 2010;**81**:257–282.