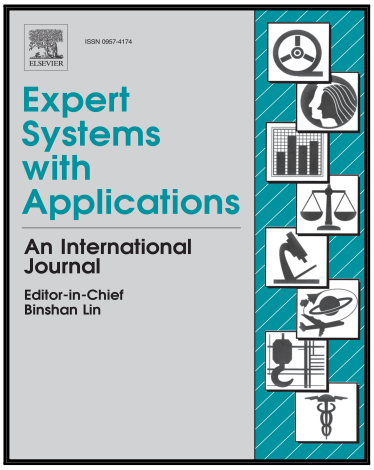
Journal Pre-proof



Model validation failure in class imbalance problems

Seokho Kang

PII: S0957-4174(20)30016-6
DOI: https://doi.org/10.1016/j.eswa.2020.113190
Reference: ESWA 113190

To appear in: *Expert Systems With Applications*

Received date: 13 February 2019
Revised date: 6 August 2019
Accepted date: 6 January 2020

Please cite this article as: Seokho Kang, Model validation failure in class imbalance problems, *Expert Systems With Applications* (2020), doi: https://doi.org/10.1016/j.eswa.2020.113190

Highlights

- Model validation is inherently difficult under class imbalance where minority class is rare in absolute sense
- Validation performance would misrepresent generalization ability of classification models
- Random guessing models can yield considerably high validation performance by chance
- Higher degree of absolute rarity contributes to increased likelihood of model validation failure

Model validation failure in class imbalance problems

Seokho Kang[a,*]

[a]*Department of Systems Management Engineering, Sungkyunkwan University, 2066 Seobu-ro, Jangan-gu, Suwon 16419, Republic of Korea*

Abstract

For a classification task, multiple classification models can be built from the training set in various ways. In general, the best-performing model is selected for deployment through a model validation procedure. However, even if the dataset is sufficiently large, model validation is difficult when the minority class is too rare in an absolute sense in the validation set. Under such an extreme absolute rarity condition, the validation performance of a model is more affected by randomness in the model so that it would misleadingly estimate the generalization ability of the model. In this regard, even a random guessing model, which will eventually fail to accurately classify new data, can yield a considerably high validation performance by chance. This implies that the selected model may not perform well during its deployment. In this study, the effect of absolute rarity on the inherent difficulty of model validation is investigated. We demonstrate that the higher degree of absolute rarity in the validation set as well as comparing a larger number of models during model validation contribute to an increased likelihood of model validation failure. Finally, a practical guideline is suggested to evaluate model validation results.

Keywords: class imbalance, model validation, absolute rarity, performance evaluation

1. Introduction

The class imbalance problem, where a dataset exhibits an unequal distribution between its classes, is prevalent in real-world classification tasks, such as fault diagnosis, fraud detection, spam filtering, and response modeling. The minority class of an imbalanced dataset is usually related to the main interest of a target task. Thus, it substantially involves greater cost when its instances are incorrectly classified. However, conventional learning algorithms tend to perform poorly in the minority class because they generally presume a well-balanced class distribution, which often makes the practical classification model trained on an imbalanced dataset biased toward the majority class. The performance of the model can further get degraded when class imbalance is severe (Lee & Cho, 2006). Dealing with class imbalance has been one of the major challenges and has received much attention in the data mining research society (Yang & Wu, 2006).

Given a classification task with an imbalanced dataset, a well-performing classification model is obtained according to the general modeling procedure as follows (Jain et al., 2000). Firstly, the original dataset is divided into three disjoint subsets: training, validation, and test sets. Several candidate models are built on the training set by employing various learning algorithms with different preprocessing, dimensionality reduction, and hyperparameter

*Corresponding author. Tel.: +82 31 290 7596; Fax: +82 31 290 7610.
Email address: s.kang@skku.edu (Seokho Kang)

settings. The classification performance of each model is then evaluated on the validation set, which consists of instances that have never been observed during training, using the selected performance measure. The model with the best validation performance is finally evaluated on the test set to analyze its generalization ability.

To address class imbalance in the training set, there have been considerable research efforts on devising various methodologies to obtain more practical classification models (He & Garcia, 2009; Japkowicz & Stephen, 2002; López et al., 2013; Sun et al., 2009). The most common approach is the use of resampling techniques (*e.g.*, oversampling and undersampling) as a preprocessing step in order to modify the class distribution of the training set to become more balanced, thereby allowing conventional learning algorithms to perform well (Estabrooks et al., 2004; Chawla et al., 2002; Liu et al., 2009; Batista et al., 2004; Maldonado et al., 2019; Zhu et al., 2019; Cateni et al., 2014). Another popular approach is cost-sensitive learning, which assigns higher cost on misclassifying minority class instances at the algorithmic level (Domingos, 1999; Elkan, 2001; Iranmehr et al., 2019). Some studies examined the effectiveness of one-class classification models to be applied to the imbalanced training set (Raskutti & Kowalczyk, 2004; Lee & Cho, 2006; Japkowicz, 2001). More recently, ensemble learning-based (Fernandes & de Carvalho, 2019; He et al., 2018), feature learning-based (Liu et al., 2018; Bellinger et al., 2018), transfer learning-based (Al-Stouhi & Reddy, 2016), and generative learning-based approaches (Douzas & Bacao, 2018; Fiore et al., 2019) have been presented. They have been successful in handling class imbalance in the training set.

Regarding performance evaluation under class imbalance in the validation set, the selection of a proper performance measure is an important aspect for model validation. Classification accuracy, which denotes the fraction of the correctly classified instances, is inadequate despite its popularity because it assumes an equal misclassification cost for all classes. Thus, the naïve approach of classifying every instance as the majority class provides near-perfect accuracy under severe class imbalance. Therefore, several alternative measures, such as precision/recall, $F1$-score, and G-mean, have been employed to overcome the limitations of classification accuracy. These measures can also be used to optimize the decision threshold of a classification model. In addition, for more reliable performance evaluation, threshold-independent measures that unify various possible settings of the decision threshold, such as a receiver operating characteristics (ROC) curve (Fawcett, 2006) and precision-recall curve (Davis & Goadrich, 2006), have been widely used in practice. However, when the minority class is too rare in an absolute sense in the validation set, model validation would sometimes fail to select a superior model that generalizes well to the test set.

This study investigates the issue of model validation failure under class imbalance, focusing on absolute rarity of the highly imbalanced validation set. With absolute rarity, a random guessing model, which will eventually fail to classify unseen instances accurately, can yield a considerably high validation performance by chance. This implies that, even if the input variables do not provide any meaningful information to predict the output variable, some classification models can yield high classification performance by chance during model validation. Accordingly, the real capabilities of classification models would not be evaluated adequately, resulting in misleading model validation results. Inferior models, which are more likely to fail to make accurate prediction for future instances, can be selected by erroneous performance evaluation. An empirical study is conducted to demonstrate model validation failure under various class imbalance conditions.

The remainder of this paper is organized as follows. In Section 2, we present the model validation and selection procedures under class imbalance scenarios. In Section 3, the effect of absolute rarity on model validation failure is

		Predicted	
		Positive	Negative
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	False Positive (FP)

Fig. 1: Confusion matix.

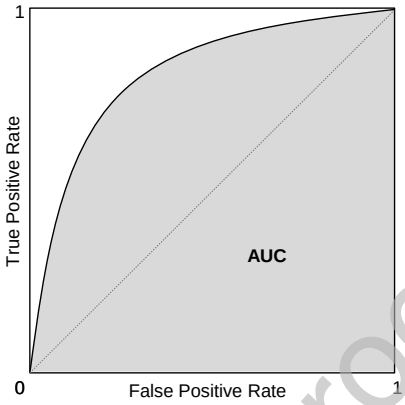


Fig. 2: Concept of ROC curve and AUC.

examined. In Section 4, the results and discussion on the empirical study are reported. Finally, the conclusions of this study are presented in Section 5.

2. Model validation and selection

In this study, we concentrate on two-class classification tasks in which many instances for the majority class and only a few instances for the minority class are available in the validation set. As the performance measure, we assume an the area under the ROC curve (AUC), which has proven to be reliable for the class imbalance problem (Huang & Ling, 2005; Bradley, 1997; Fawcett, 2006). The ROC curve plots the trade-off between the true positive (TP) and the false positive (FP) rates with varying decision thresholds of a classification model. The TP and FP rates for a given decision threshold are calculated as TP/(TP+FN) and FP/(FP+TN), respectively, based on the confusion matrix shown in Fig. 1. The AUC stands for the area under the curve. Fig. 2 shows an example of an ROC curve and the corresponding AUC. AUC is calculated independent of the decision threshold by integrating the performance of the model over all possible settings of the threshold. A larger AUC indicates a better trade-off between the TP and FP rates. A perfect classification model yields 1 for the AUC. A classification model that satisfies AUC>0.5 is considered acceptable; otherwise, it is non-informative thereby making them negligible.

For a classification task, model validation is essential to ensure the generalization ability of a classification model. During model validation, the classification performance of the model is evaluated on the validation set, which consists of instances that have not been used for its training to prevent overfitting. Once multiple candidate models are obtained, the model demonstrating the highest validation performance is selected to be used for classifying unseen instances in the future. Therefore, the success of model deployment highly depends on the adequacy of performance

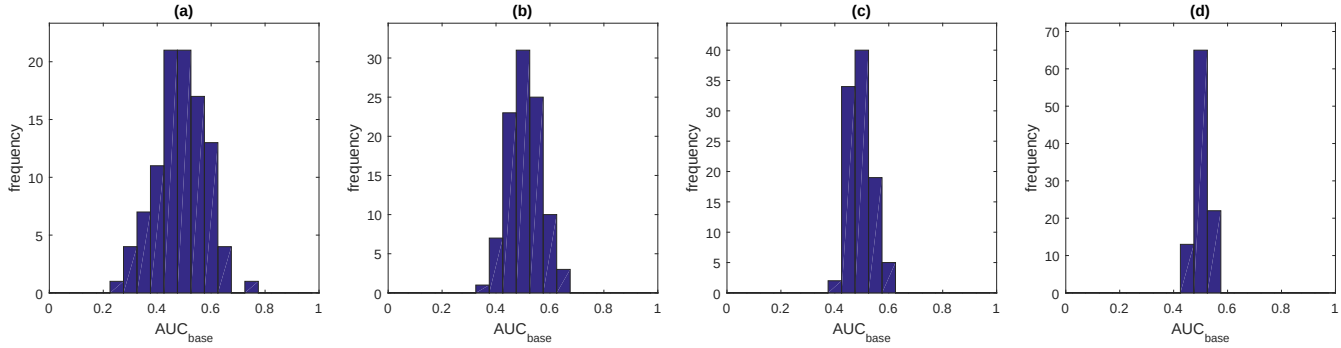


Fig. 3: $\mathtt{AUC}_{\mathrm{base}}$ distribution obtained by random guessing model. N=10000, m = (a) 10, (b) 20 (c) 50, and (d) 100.

evaluation during model validation.

The origin of the performance of a model can be described in two parts. A substantial portion of the performance is explained as the reflection of underlying information in data, while another part can be obtained by chance independent of data. In terms of the AUC, the performance of a classification model can be described as follows, using two elements $\mathtt{AUC}_{\mathrm{base}}$ and $\mathtt{AUC}_{\mathrm{gain}}$:

$$\mathrm{AUC} = \mathtt{AUC}_{\mathrm{base}} + \mathtt{AUC}_{\mathrm{gain}}. \tag{1}$$

It should be noted that the AUC cannot be directly decomposed into $\mathtt{AUC}_{\mathrm{base}}$ and $\mathtt{AUC}_{\mathrm{gain}}$ for a model.

$\mathtt{AUC}_{\mathrm{base}}$ is the baseline of the AUC that can be simply obtained based on its own property without using any information in data. A random guessing model, which simply assigns an arbitrary score as the posterior probability for each instance, is expected to yield the AUC value of 0.5. It can be considered the baseline. In this study, we use the AUC of the random guessing model as the estimation of $\mathtt{AUC}_{\mathrm{base}}$. $\mathtt{AUC}_{\mathrm{base}}$ is determined without learning from data, thus depends inherently on randomness. Fig. 3 shows the histograms of the estimated $\mathtt{AUC}_{\mathrm{base}}$ distributions, each of which is obtained using 100 independent random guessing models, for four different conditions of class imbalance with respect to the minority class size m while the validation set size N is fixed at 10000. Each distribution is centered near 0.5, and is more dispersed when the minority class is rarer. An ROC curve becomes less smooth with a smaller m, which is another reason that causes the high variation in an $\mathtt{AUC}_{\mathrm{base}}$ distribution.

In contrast, $\mathtt{AUC}_{\mathrm{gain}}$ is the improvement of the AUC against the baseline by exploiting the information in data. While a random guessing model provides an $\mathtt{AUC}_{\mathrm{gain}}$ value of 0, a classification model that learned to capture the underlying relationships between the input and output variables would yield an $\mathtt{AUC}_{\mathrm{gain}}$ value that is greater than 0. In addition, $\mathtt{AUC}_{\mathrm{gain}}$ becomes higher when the input variables are more informative to predict the output variable. $\mathtt{AUC}_{\mathrm{gain}}$ is closely related to the generalization ability of a classification model. If the model overfits the training set, $\mathtt{AUC}_{\mathrm{gain}}$ obtained during model validation is likely to be lower than that during model training. Therefore, it is more important to take $\mathtt{AUC}_{\mathrm{gain}}$ into account during model validation to properly select the model with superior generalization ability.

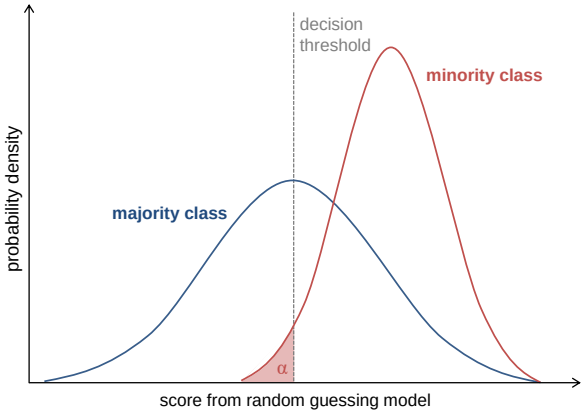


Fig. 4: Distribution bias of minority class in validation set.

3. Model validation failure under class imbalance

The class imbalance problem originates from the rarity of the minority class in data. Two types of rarity can be considered: relative rarity and absolute rarity (He & Garcia, 2009; Weiss, 2004). Relative rarity is associated with when the minority class instances are relatively rarer compared to the other class, *i.e.*, the value of m/N is very small. On the other hand, absolute rarity describes the lack of minority class instances in an absolute sense, *i.e.*, the value of m is very small, which is the main focus of this work. Let us suppose that the imbalance ratio of the validation set is extremely high (*e.g.*, N:m = 1000:1, 10000:1, or more). If the validation set size N is very large, a sufficient number of instances that represent the minority class can still be available for model validation. Thus, well-performing classification models can be obtained by employing certain methodologies of dealing with class imbalance. However, absolute rarity introduces the risks that the minority class is underrepresented and the output variable is coincidentally correlated to random noise. This causes a classification model to overact to minor fluctuations in the output variable during model validation.

This study focuses particularly on the influence of extreme absolute rarity of minority class instances in the validation set on the failure of model validation. When the minority class size m is very small, some classification models can yield smaller or larger values of $\texttt{AUC}_{\text{base}}$ by chance during model validation irrespective of the data characteristics as the variability of $\texttt{AUC}_{\text{base}}$ becomes greater. The deviation of $\texttt{AUC}_{\text{base}}$ from 0.5 is likely to become larger than $\texttt{AUC}_{\text{gain}}$ so that the AUC is mostly affected by $\texttt{AUC}_{\text{base}}$ instead of $\texttt{AUC}_{\text{gain}}$. Accordingly, it is naturally difficult to select the real best model through model validation when comparing multiple candidate models.

We would like to demonstrate that, under extreme absolute rarity, even a random guessing model whose expected AUC value on the test set is 0.5 can achieve a high AUC value during model validation. For a simplified illustration, it is assumed that the entire validation set is split into two disjoint equal-sized subsets using a decision threshold cutoff for a random guessing model, as shown in Fig. 4. The majority class is evenly distributed over the two subsets, whereas the minority class is more likely to be biased toward one side of the two subsets when the minority class is absolutely rare, *i.e.*, the distribution of the minority class is biased. Assuming that the majority class is much larger than the minority class (*i.e.*, $N >> m$), the estimated probability that more than 1-α fraction of the minority class instances are on one side is $\text{p}(m, \alpha) = \sum_{k \leq m \cdot \alpha} \binom{m}{k} / 2^{m-1}$. A smaller α indicates a greater threshold for the

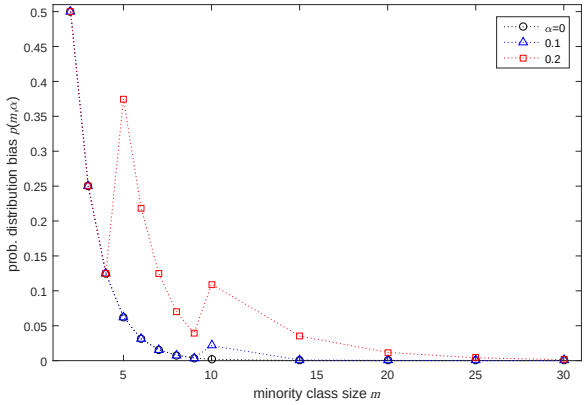


Fig. 5: Probability of minority class distribution being biased with varying m and α.

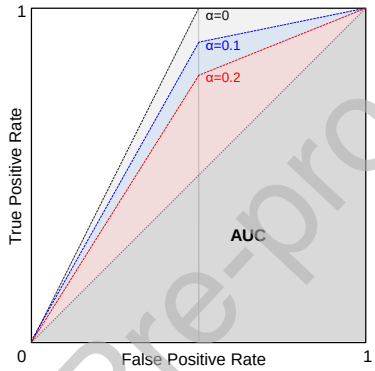


Fig. 6: Expected ROC curve when minority class distribution is biased with α=0, 0.1, and 0.2.

distribution bias of the minority class, and therefore, the other side contains less than α fraction of the minority class instances. When $\alpha = 0$, all minority class instances belong to one side, and the corresponding probability is approximately $\mathrm{p}(m, 0) = \binom{m}{0}/2^{m-1}$.

Fig. 5 plots the values of $p(m, \alpha)$ calculated by varying the minority class size m and distribution bias threshold α. As shown in the figure, $p(m, \alpha)$ becomes larger as the degree of absolute rarity increases with smaller m and as a larger value is assigned to α. If we have $1/p(m, \alpha)$ different random guessing models, one model is expected to have a biased distribution with respect to α on the minority class. For example with m=20 and α=0.2, one out of $1/p(20, 0.2) \simeq 84.6$ random guessing models is expected to have a biased minority class distribution of α=0.2. The random guessing models exhibiting greater distribution bias of the minority class would yield higher AUC values. Fig. 6 illustrates the expected ROC curve on a random guessing model with α=0, 0.1, and 0.2. For α=0, the expected value of AUC is 0.75, which is a considerably high performance in practice (Swets, 1988). The expected values of AUC are 0.7 and 0.65 regarding α=0.1 and 0.2, respectively, which are not negligible as well. Although all random guessing models will eventually fail to classify unseen instances accurately, some of them are more likely to yield high AUC values by chance in the model validation procedure.

This simple illustration demonstrates that even a random guessing model can achieve a considerably high AUC by chance owing to the high variability of $\mathtt{AUC}_{\mathrm{base}}$ for the imbalanced validation set with absolute rarity. Given

Table 1: Experimental settings for model validation scenarios.

Scenario	Description	Conditions of class imbalance			
		Size of validation set (N)	Size of minority class (m)	Fraction of minority class (m/N)	Number of models compared (l)
Scenario 1	Fixed validation set size, Effect of absolute rarity	10000 (fixed)	10, 20, 50, 100	0.1%, 0.2%, 0.5%, 1%	10, 20, 50, 100, 200, 500
Scenario 2	Fixed relative rarity, Effect of absolute rarity	1000, 2000, 5000, 10000	10, 20, 50, 100	1% (fixed)	
Scenario 3	Fixed absolute rarity, Effect of relative rarity	1000, 2000, 5000, 10000	50 (fixed)	5%, 2.5%, 1%, 0.5%	

the same training set, different candidate classification models are obtained using different ways of modeling. For example, complex learning algorithms involving several hyperparameters (*e.g.*, support vector machines and artificial neural networks) and wrapper feature selection methods (*e.g.*, sequential forward/backward selection and genetic algorithm) require a huge number of candidate models to be compared during model validation. Each model has a different value for $\texttt{AUC}_{\text{base}}$ on the validation set. When more candidate models are compared, some of them are more likely to achieve higher $\texttt{AUC}_{\text{base}}$ values on the validation set. Model validation would be successful if some models succeed in yielding a sufficiently high value of $\texttt{AUC}_{\text{gain}}$ as compared to the variability of $\texttt{AUC}_{\text{base}}$. Otherwise, the AUC of each model on the validation set is determined by randomness in the model, which could make the results of model validation misleading.

4. Empirical analysis

An empirical analysis was conducted to examine the class imbalance factors that affect the variability of $\texttt{AUC}_{\text{base}}$ more during the model validation procedure, thereby causing model validation failure. To evaluate the degree of model validation failure, we demonstrated the best validation performance that can be attained by simply comparing a number of random guessing models independent of input variables under different class imbalance conditions. $\texttt{AUC}_{\text{gain}}$, which is obtained by identifying the relationship between input and output variables in data, was not included in the analysis.

We considered the following three model validation scenarios, which pertain to different class imbalance factors in the output variable in the validation set. In scenarios 1 and 2, the effect of absolute rarity was demonstrated when the validation set size N and degree of relative rarity (m/N) were fixed, respectively. Scenario 3 dealt with the effect of relative rarity on the same degree of absolute rarity. The effect of the number of models compared during model validation (l) was also investigated for each scenario. The conditions for class imbalance in terms of N, m, and l used for each scenario are listed in Table 1.

For each scenario, we compared the classification performance on those conditions with respect to N, m, and l in terms of $\texttt{AUC}_{\text{base}}$ obtained by validating multiple random guessing models and taking the AUC from the best

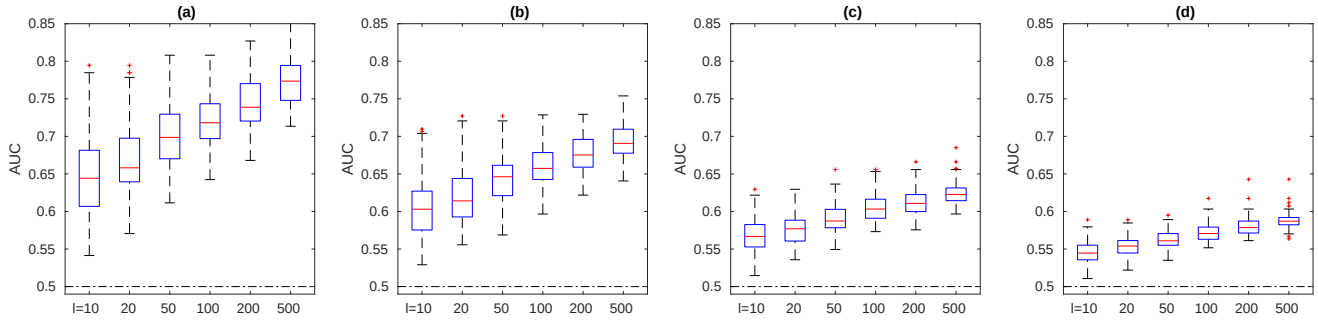


Fig. 7: Comparison results of AUC distributions for scenario 1: (a) m=10, (b) 20, (c) 50, and (d) 100; N = 10000.

one. In detail, the best $\mathtt{AUC}_{\mathrm{base}}$ for a given specific condition was estimated by conducting the following modeling procedure. For each condition, we firstly generated an artificial validation set of size N with one output variable, for which only m instances were labeled as the minority class. It should be noted that input variables are not considered because we evaluated the performance of random guessing models that make predictions independently of input variables. Then, l different random guessing models were evaluated in terms of the AUC on the validation set. The highest value of the AUC across these l models was finally selected to be used as the representative value for the condition. With 100 independent runs of this procedure, the resulting set of the values constituted the estimated $\mathtt{AUC}_{\mathrm{base}}$ distribution of the condition. Thus, each value in the distribution corresponds to the best random guessing model during model validation at an independent run. All implementations were conducted using MATLAB.

For scenario 1, Fig. 7 shows the box plots of the estimated $\mathtt{AUC}_{\mathrm{base}}$ distributions on different class imbalance conditions with respect to m and l for the same N. In each box plot, the box and whisker represent the interquartile range and min-max range, respectively. The line inside the box indicates the median value. As shown, both values and the deviation of distribution tended to increase with a higher degree of absolute rarity, which indicates a greater chance for a model to achieve a higher value of $\mathtt{AUC}_{\mathrm{base}}$. The values of $\mathtt{AUC}_{\mathrm{base}}$ also tended to increase when a larger set of models were compared during each independent run, because this set was more likely to contain the models with higher $\mathtt{AUC}_{\mathrm{base}}$. For the condition of l=500 and m=10, the maximum and median values in the distribution were 0.861 and 0.774, respectively, which are regarded as considerably high. Even the minimum value in the distribution regarding this condition was 0.714, which is much larger than the general expectation of the AUC for a single random guessing model. On the other hand, for the condition of l=10 and m=100, the distribution was much closer to the baseline value of 0.5.

The box plots in Fig. 8 compare the estimated $\mathtt{AUC}_{\mathrm{base}}$ distributions for conditions in scenario 2 with varying m and N while m/N is fixed at 1%. Similar to scenario 1, the values and the deviation of the distribution tended to increase with a higher degree of absolute rarity for the same degree of relative rarity. For the conditions with l=500, the median value in the distribution was 0.765 when m=10, but decreased to 0.587 when m=100. In addition, the distribution shifted upward as l became larger.

The comparison results for scenarios 1 and 2 demonstrated that, either for the same validation set size or same degree of relative rarity, absolute rarity was closely related to the $\mathtt{AUC}_{\mathrm{base}}$ distributions. The higher degree

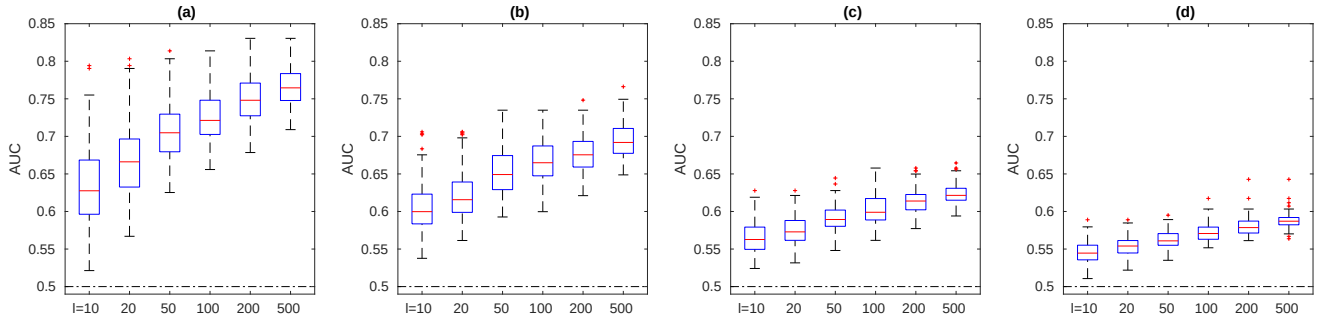


Fig. 8: Comparison results of AUC distributions for scenario 2: (a) (m,N)=(10,1000), (b) (20,2000), (c) (50,5000), and (d) (100,10000).

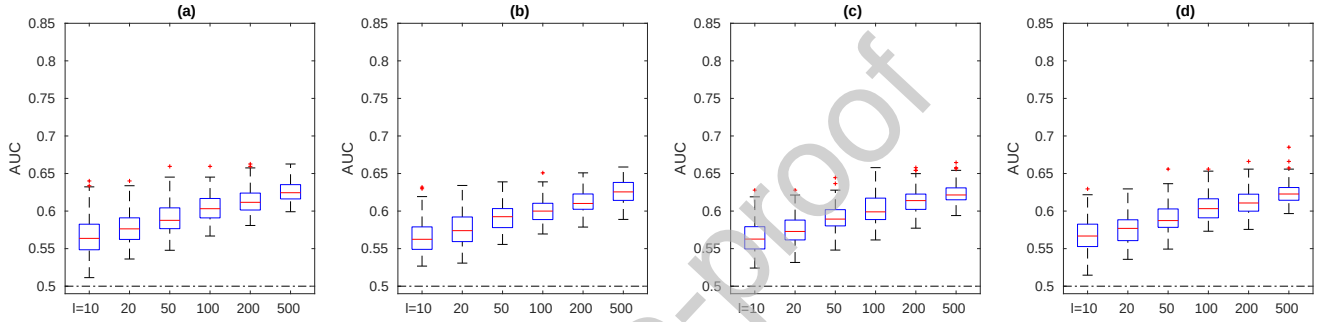


Fig. 9: Comparison results of AUC distributions for scenario 3: m=50; (a) N=1000, (b) 2000, (c) 5000, and (d) 10000.

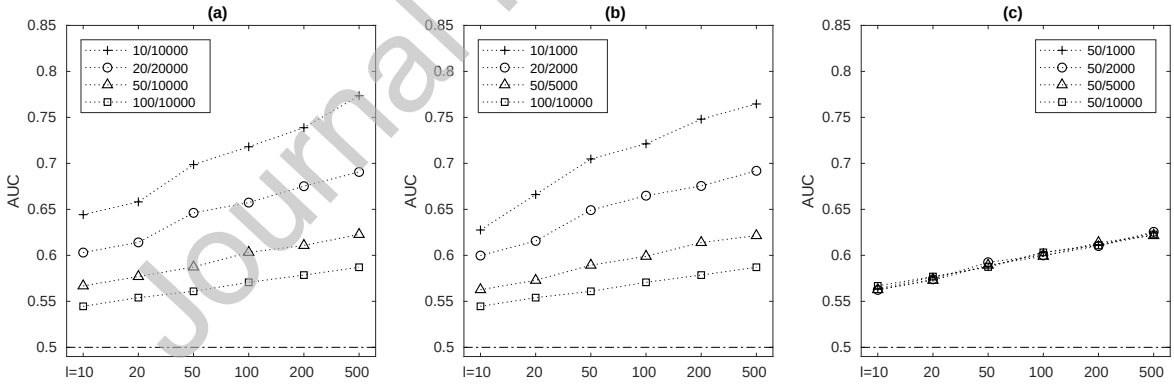


Fig. 10: Overall comparison results for (a) scenario 1, (b) scenario 2, and (c) scenario 3.

of absolute rarity digressed the distributions from 0.5 because performance evaluation of classification models was more likely to involve randomness, thereby providing misleading model validation results.

Regarding scenario 3, the different degrees of relative rarity on the same degree of absolute rarity are compared in Fig. 9. As shown in the results, no significant difference was observed between the $\mathtt{AUC}_{\mathrm{base}}$ distributions of the different degrees of relative rarity. This infers that relative rarity was not a significant factor that affected the distributions.

Fig. 10 illustrated the overall comparison on the median values of the $\texttt{AUC}_{\text{base}}$ distributions from Figs. 7, 8, and 9. To summarize, it was demonstrated that two main factors yielded values of $\texttt{AUC}_{\text{base}}$ that were much higher than 0.5. First, a higher $\texttt{AUC}_{\text{base}}$ value was obtained with by a greater degree of absolute rarity. Second, $\texttt{AUC}_{\text{base}}$ increased further when more candidate classification models were compared during model validation. Owing to the higher variability of $\texttt{AUC}_{\text{base}}$ in such extreme conditions, the $\texttt{AUC}_{\text{gain}}$ of a particular model that learned from data would not be a major factor in the determination of the AUC of the model when comparing them. Moreover, even a random guessing model would yield higher AUC based on the variability of $\texttt{AUC}_{\text{base}}$. In this regard, the generalization ability of these models cannot be properly evaluated during model validation, thereby resulting in inadequate model selection results. Accordingly, it is concluded that the validation set exhibiting higher degree of absolute rarity would be subjected to greater potential of model validation failure.

There is no definite solution to avoid model validation failure in classification tasks due to absolute rarity. Instead, we can suggest a practical guideline to evaluate the confidence of model validation results. Given a classification task with a highly imbalanced validation set exhibiting absolute rarity, if the best AUC obtained during model validation by comparing multiple classification models that learned from data is significantly larger than the estimated $\texttt{AUC}_{\text{base}}$ distribution with respect to the equivalent condition of N, m, and l, the $\texttt{AUC}_{\text{gain}}$ of the selected best model is sufficiently large to ensure the generalization ability of the model. Otherwise, we cannot guarantee that the best model is well-performing in terms of generalization ability. For example, let us assume a model validation procedure comparing the AUC of 100 different classification models on the validation set of N=10000 and m=50. Then, the estimated 95% confidence interval of the highest $\texttt{AUC}_{\text{base}}$ obtained from the procedure would be $(0.589, 0.658)$ by assuming a t-distribution. The best model would be meaningful and useful if the AUC of the model is much larger than the interval, whereas it is difficult to determine whether the model is reliable in its deployment if the AUC is not sufficiently large.

5. Conclusion

In this study, we investigated model validation failure for highly imbalanced data exhibiting absolute rarity, and performed an empirical study to demonstrate it on diverse conditions of class imbalance. As the AUC is more influenced by the randomness, even a random guessing model can even yield considerably high AUC by chance owing to its high variability. This indicates that the AUC cannot properly represent the generalization ability of a classification model, *i.e.*, the variability of $\texttt{AUC}_{\text{base}}$ is large enough to discard the effect of $\texttt{AUC}_{\text{gain}}$. Therefore, even if we adopt sophisticated methodologies to obtain well-performing classification models, model validation failure would be inherently inevitable. This implies that it is likely to fail to select the best model when comparing multiple models during model validation. Moreover, a higher degree of absolute rarity gives rise to a greater likelihood of model validation failure.

Our empirical study showed the estimated $\texttt{AUC}_{\text{base}}$ distributions for different conditions of class imbalance, which can be used as a guideline to assess the confidence of model validation regarding a classification task that involves the class imbalance problem. The model validation procedure would be considered successful if the procedure for the validation set provides a much higher value of the AUC as compared to the estimated $\texttt{AUC}_{\text{base}}$ distribution of the

corresponding condition. Otherwise, the results of the procedure would be less confident, and we cannot guarantee that the best model selected by the procedure will perform well during its deployment.

We believe this study serves as a guideline for both researchers and practitioners to properly address class imbalance. To circumvent the problem of model validation failure, it is important to lower the effect of randomness for the evaluation of validation performance of a model to analyze the generalization ability of the model. It would be better to start by evaluating only a few models that are simple and robust for model validation, which allows the corresponding $\texttt{AUC}_{\text{base}}$ distribution to be closer to 0.5 by reducing the randomness. In addition, an attempt should be made first to secure more minority class instances to mitigate the degree of absolute rarity.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT; Ministry of Science and ICT) (No. NRF-2017R1C1B5075685).

References

References

Al-Stouhi, S., & Reddy, C. K. (2016). Transfer learning for class imbalance problems with inadequate data. *Knowledge and Information Systems*, *48*, 201–228.

Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*, 20–29.

Bellinger, C., Drummond, C., & Japkowicz, N. (2018). Manifold-based synthetic oversampling with manifold conformance estimation. *Machine Learning*, *107*, 605–637.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*, 1145–1159.

Cateni, S., Colla, V., & Vannucci, M. (2014). A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, *135*, 32–41.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of International Conference on Machine Learning* (pp. 233–240).

Domingos, P. (1999). MetaCost: a general method for making classifiers cost-sensitive. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 155–164).

Douzas, G., & Bacao, F. (2018). Effective data generation for imbalanced learning using conditional generative adversarial networks. *Expert Systems with Applications*, *91*, 464–471.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of International Joint Conference on Artificial Intelligence* (pp. 973–978).

Estabrooks, A., Jo, T., & Japkowicz, N. (2004). A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, *20*, 18–36.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*, 861–874.

Fernandes, E. R. Q., & de Carvalho, A. C. P. L. F. (2019). Evolutionary inversion of class distribution in overlapping areas for multi-class imbalanced learning. *Information Sciences*, *494*, 141–154.

Fiore, U., De Santis, A., Perla, F., Zanetti, P., & Palmieri, F. (2019). Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. *Information Sciences*, *479*, 448 – 455.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284.

He, H., Zhang, W., & Zhang, S. (2018). A novel ensemble method for credit scoring: adaption of different imbalance ratios. *Expert Systems with Applications*, *98*, 105–117.

Huang, J., & Ling, C. X. (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, *17*, 299–310.

Iranmehr, A., Masnadi-Shirazi, H., & Vasconcelos, N. (2019). Cost-sensitive support vector machines. *Neurocomputing*, *343*, 50–64.

Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 4–37.

Japkowicz, N. (2001). Supervised versus unsupervised binary-learning by feedforward neural networks. *Machine Learning*, *42*, 97–122.

Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: a systematic study. *Intelligent Data Analysis*, *6*, 429–449.

Lee, H.-j., & Cho, S. (2006). The novelty detection approach for different degrees of class imbalance. In *Proceedings of International Conference on Neural Information Processing* (pp. 21–30).

Liu, X.-Y., Wu, J., & Zhou, Z.-H. (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*, 539–550.

Liu, Y.-H., Liu, C.-L., & Tseng, S.-M. (2018). Deep discriminative features learning and sampling for imbalanced data problem. In *Proceedings of IEEE International Conference on Data Mining* (pp. 1146–1151).

López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, *250*, 113–141.

Maldonado, S., López, J., & Vairetti, C. (2019). An alternative SMOTE oversampling strategy for high-dimensional datasets. *Applied Soft Computing*, *76*, 380–389.

Raskutti, B., & Kowalczyk, A. (2004). Extreme re-balancing for SVMs: a case study. *ACM SIGKDD Explorations Newsletter*, *6*, 60–69.

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: a review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*, 687–719.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Weiss, G. M. (2004). Mining with rarity: a unifying framework. *ACM SIGKDD Explorations Newsletter*, *6*, 7–19.

Yang, Q., & Wu, X. (2006). 10 challenging problems in data mining research. *International Journal of Information Technology & Decision Making*, *5*, 597–604.

Zhu, T., Lin, Y., Liu, Y., Zhang, W., & Zhang, J. (2019). Minority oversampling for imbalanced ordinal regression. *Knowledge-Based Systems*, *166*, 140–155.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.