MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Research Article

Open Access

CrossMark

Classification and prediction with very imbalanced group sample sizes: an illustration with COVID-19 testing

Abstract

This study explored predictions of COVID test results using statistical classification methods based on available COVID-related data such as demographic and symptom information. The performances of logistic regression, machine learning models, and latent class analysis in the predictions of extreme imbalanced COVID data were compared. One technical challenge of using statistical classification methods was tackled in the extreme imbalance sample sizes of the COVID data. The oversampling method was applied on the training dataset to mitigate the impact of such data structure on the training process. Further, the adjusted pooled sampling method based on the statistical classification results was proposed to facilitate the efficiency of COVID testing. Results indicate that some machine learning models (e.g., support vector machine) had better performance than traditional logistic regression model and latent class analysis under extreme imbalance data condition. Further, the oversampling method increased the sensitivity of various statistical classification methods when different cut-off values were applied. The adjusted pooled sampling was shown to be more efficient than the traditional pooled sampling method.

Keywords: COVID-19 data, machine learning, logistic regression, latent class analysis, unbalanced samples

Volume 10 Issue 4 - 2021

Xin Qiao, Yishan Ding, Chengbin Ying, Hong Jiao, George Macready

Human Development & Quantitative Methodology, University of Maryland, College Park, USA

Correspondence: Hong Jiao, Human Development & Quantitative Methodology, University of Maryland College Park, USA, Tel 3019685138, Email hjiao@umd.edu

Received: October 10, 2021 | **Published:** November 19, 2021

Introduction

The outbreak of the COVID-19 pandemic has brought numerous challenges to the healthcare system worldwide. A timely and comprehensive detection of carriers is a crucial strategy employed by many countries to control the virus's spread to a relatively successful degree. However, the virus and antibody's medical test is both time-consuming and expensive, and the cost grows proportionally with population size. Alternative methods that rely on the demographic information and self-reported symptoms are thus promising to serve as a preliminary screen model to predict individual-level risks in COVID-positivity. Statistical analyses of such data not only provide risk evaluation about individual virus infection, but also are conducive for refined medical testing procedures.

Since the inception of the disease, the most frequently used prediction model in the analysis of COVID data is the logistic regression model).[1-4] In a systematic review study, Waynants et al.[4] reported 107 studies describing 145 prediction models for diagnosis and prognosis of COVID-19, with 4 models predicting risk of hospital admission for COVID-19 pneumonia in the general population, 91 diagnostic models for detecting COVID-19 patients with suspected infection, and 50 prognostic models for predicting mortality risk, progression to severe disease, or length of hospital stay. A majority of these studies focused on the significance of the predictors in their relationship with the outcome variable (i.e., positive or negative cases) using the logistic regression model. The most frequently reported predictors of the presence of COVID-19 included age, body temperature, signs and symptoms, sex, blood pressure, and creatinine. Marin et al.,[3] similarly, summarized factors predictive of increased disease severity and/or mortality as the following: age > 55 years, multiple pre-existing comorbidities, hypoxia, specific computed tomography findings indicative of extensive lung involvement, diverse laboratory test abnormalities, and biomarkers of end-organ dysfunction. C index estimates, i.e., summaries of discrimination quantifying the extent to which predicted risks discriminate between participants with and without the outcome for these models, ranged from 0.73 to 0.81 in prediction models for the general population, from 0.65 to 0.99 in diagnostic models, and from 0.85 to 0.99 in prognostic models, respectively.

In addition to traditional logistic regression, several studies include procedures for variable selection. Jehi et al.[2] developed the least absolute shrinkage and selection operator logistic regression algorithm to retain variables that contribute to the model prediction. The model focused on predicting the likelihood of a positive nasal or oropharyngeal COVID-19 test, using a prospective registry of all patients tested for COVID-19 in the Cleveland Clinic. Results of the study suggested that male, African American, older patients, and those with known COVID-19 exposure were at higher risk of being positive for COVID-19. In comparison, risk was reduced in patients who had pneumococcal poly-saccharide or influenza vaccine or who were on melatonin, paroxetine, or carvedilol. C-statistic was 0.863 in the development cohort and 0.840 in the validation cohort, which indicates similar prediction accuracy as the traditional logistic regression reported in Marin et al.[3] Similarly, Sun et al.[5] developed four multivariate logistic regression models predicting positive COVID-19 cases using variables selected through stepwise use of Akaike's information criterion (AIC; Akaike,[6]), while Bhargava et al.[1] incorporated the forward likelihood ratio algorithm to build multivariate logistic regression models. Results of the Sun et al.[5] study suggested that positive cases were more likely to be older people with comorbidities, those that had contact with a known COVID-19 case or had recently traveled to China. Further, these cases were more likely to have an elevated body temperature at clinical presentation and radiological findings suggestive of pneumonia as well as lower blood counts of white blood cells, platelets, neutrophils, lymphocytes, eosinophils, and basophils. The predictive performance of the final models was assessed using receiver operating characteristic (ROC)

curves and the corresponding area under the curve (AUC) values. All models had satisfactory prediction results with AUCs of 0.91, 0.88, 0.88 and 0.65 for the four models, respectively. Bhargava et al.[1] reported similar results in a study designed to identify predictors for severe COVID-19 infection. The four significant predictors for severity of COVID-19 infection included presence of pre-existing renal disease, need for supplemental oxygen at the time of hospitalization, elevated creatinine and C-reactive protein (CRP) on admission laboratory findings.

In addition to the prevalent use of multivariate logistic regression models on the COVID-19 data, machine learning algorithms have been used to build prediction models for efficient diagnosis purposes.[4] Kumar et al.,[7] for example, used random forest and XGBoost predictive classifiers on chest X-ray images and found prediction accuracies as 0.973 and 0.977, respectively. Similarly, Hassanien et al.[8] used support vector machine to detect COVID cases using the lung X-ray images and found accuracies higher than 0.950. These studies show the potential of machine learning methods in the early detection of COVID-19.

One technical issue in the analysis of COVID data is the extreme imbalance of the outcome variable (i.e., positive vs. negative cases). COVID data is a typical type of "rare events data" in that the binary outcome variables contain far more negative cases than positive cases.[9] The large discrepancy between positive and negative cases suggests that the predictive model should not only be evaluated based on total classification accuracy because a model that classifies all individuals into negative cases can yield a high accuracy as well. The model diagnostic and comparison should also consider sensitivity, computed as the proportion of correct detection of all positive cases, to be the primary evaluation criterion. Further, the impact of the imbalance in sample sizes in the COVID prediction model also calls for statistical adjustment that takes into account the rare events. For example, Kumar et al.[7] used the resampling method to balance the positive and negative data points. However, few studies have compared the performances of classifiers using the extreme imbalanced dataset and balanced dataset.

In addition to prediction accuracy, another important issue to consider is the efficiency of COVID tests. Regular pooled sampling is conducted by testing multiple nasopharyngeal swabs simultaneously[11] to improve the testing efficiency. If the pooled test result is negative, then individuals in the pool are considered as negative cases. If the pooled test result is positive, then each individual in the pool needs to be tested again. The benefit of pooled testing is to reduce the time and financial cost of COVID testing. However, there is little guidance on how the pools should be formed. The current study proposes the adjusted pooled sampling that uses the statistical models as preliminary screening techniques utilized in combination with pooled sampling to improve the efficiency of the COVID-19 diagnosis. Specifically, the pools are formed based on the statistical prediction results. Individuals who are predicted to be positive are tested individually, while those who are predicted to be negative are put in the same pools. Given the satisfactory prediction accuracy of the statistical methods, it is expected that adjusted pooled sampling can further improve the testing efficiency.

In the current study, both the extreme imbalance issue and the pooled sampling issue are considered. Specifically, this study compares the prediction accuracy of different classification methods including logistic regression, machine learning techniques (i.e., decision tree, random forest, gradient boosting, support vector machine), and latent class analysis using empirical dataset with extreme imbalance and when the extreme imbalance is treated using resampling method. Latent class analysis has not been used in any previous studies but is considered in the current study due to its advantages for latent group classification purposes. Further, this study evaluated the performance of the adjusted pooled sampling under both the extreme imbalance scenario and when the extreme imbalance is treated. In summary, the purpose of the current study is twofold: 1) to evaluate the performances of the statistical methods in predicting positive COVID cases; 2) to examine whether the adjusted pooled sampling method can improve the efficiency of COVID testing.

Models and approaches for classification and prediction

The data analysis methods used in the current study include logistic regression, latent class analysis, and machine learning methods including classification and regression tree (CART), gradient boosting (GB), random forest (RF), and support vector machine (SVM). These methods are introduced as follows.

Logistic regression

Logistic regression model is the most frequently used method in the studies related to the COVID prediction models. It is included in the current study as a baseline method to be compared with. Let Y_j be the observed diagnosed category for person j (j = 1, ..., J) and $Y_j = 1$ indicates probable cases while $Y_j = 0$ indicates confirmed cases. In the logistic regression model, the probability of being diagnosed as a probable case is expressed as:

$$\pi(Y_j = 1) = \frac{e^{\beta^T X}}{1 + e^{\beta^T X}}, \quad (1)$$

where β denote a $p \times 1$ vector of regression coefficients and $\mathbf{X}$ denote an $n \times p$ design matrix of predictors where n is the number of observations and p is the number of predictors (including the intercept). There are several assumptions for the logistic regression model. First, it requires the dependent variable is dichotomous and sample size is large, which is the case in the current study. Second, it assumes that the observations are identically and independently distributed. Thus, the likelihood function of the logistic regression model is given by:

$$L(\pi; y_1, ..., y_j) = \prod_{j=1}^{J} \pi_j^{y_j} (1 - \pi_j)^{1 - y_j}. \quad (2)$$

The regression coefficients are estimated using the maximum likelihood estimator. Third, the logit of the probability being diagnosed as probable cases (i.e., $\pi(Y_j = 1)$) is linearly related to the predictors $\mathbf{X}$. That is,

$$\text{logit}\ [\pi(Y_j = 1)] = In\left(\frac{\pi(Y_j = 1)}{1 - \pi(Y_j = 1)}\right) = \beta^T X \quad (3)$$

Lastly, the logistic regression model assumes there is no multicollinearity issue among the predictors. That is, the predictors are not highly correlated with each other. However, the multicollinearity issue is not an issue if the goal is prediction, which is the case in the current study. The estimated regression coefficients $\hat{\beta}$ are further used in the logistic regression model to predict new observations. Therefore, all predictors were used in the logistic regression model.

Latent class analysis

The latent class analysis[11] is finite mixture modeling when the observed data are categorical. The finite mixture model has had a

long history in modeling the heterogeneous population. Everitt[12] summarizes two common scenarios when finite mixture models can help explain two types of research design: 1) the population comprises well-defined subpopulations, use the mixture model to classify individuals into the unknown classes; and 2) there are suspected subpopulations and use the mixture model for exploratory purposes. This work fits the former purpose of classifying the individuals into pre-defined subpopulations of "negative cases" versus "positive cases". LCA relies on several assumptions: 1) the data are generated from a mixture of underlying probability distributions, and the population is heterogeneous; 2) LCA assumes local independence that within each class, the endorsement of the observed variables is assumed to be mutually independent of each other.[13]

An unconditional LCA is a measurement model which specifies the relation between the measured indicators and the latent classification variable. For an individual in latent class k (k =*1, 2, ..., K*), the probability of obtaining response pattern u for the jth (j = 1, 2, …, J) indicator under the local independence assumption is given by

$$\Pr(U=u\mid c=k)=\sum_{k=1}^{K}\Pr(c=k)\prod_{j=1}^{J}\Pr(u_j=1\mid c=k), \tag{4}$$

where c is the latent class variable. Using the Bayes rule, the posterior probability of belonging to latent class k can be computed in the form

$$\Pr(c=k\mid U=u)=\frac{\Pr(U=u\mid c=k)\Pr(c=k)}{\Pr(U=u)}. \tag{5}$$

Ideally, for well-separated latent classes, all individuals have a very high model-based posterior probability of being classified into one latent class and a very low probability into the other(s).

Machine learning techniques

Four machine learning techniques are used in the current study including classification and regression tree (CART), gradient boosting (GB), random forest (RF), and support vector machine (SVM). CART, GB, and RF are tree-based methods that are fairly robust to noisy data, while SVM can handle classification issues with non-linear boundaries.[14] CART has been shown to be an effective method with simple interpretability.[15] In the current 2-class classification problem, CART divides the predictor space into m distinct and non-overlapping regions and the prediction of the classes equals the most common classes of the observations in each region. The goal is to minimize the classification error rate, which is simply the proportion of the training observations in that region that do not belong to the most commonly occurring class:

$$E=1-\max_k(\hat{p}_{ik}), \tag{6}$$

where $\hat{p}_{ik}$ indicates the proportion of training observations in the ith region that are from the kth class (k = 1 or 2 in the current study). The Gini index[16] is used to indicate the quality of a split in the current study, which is defined by

$$G=\sum_{k=1}^{K}\hat{p}_{ik}(1-\hat{p}_{ik}), \tag{7}$$

which measures the total variance across k classes. A small value of G indicates that high purity of a region which predominantly consists of observations from a single class.

However, the tree structure may be easily affected by small changes in the dataset.[17] That is, the CART method suffers from high variance in the predictions on the validation set. Therefore, RF and GB that use ensemble methods are used as comparisons to the CART method. The idea behind the ensemble methods is that many training datasets are taken from the population and the average prediction results are taken to decrease the variance in the prediction. For RF, m predictors are sampled from the full set of p predictors and one predictor from the m predictors is used for each split (usually $m\approx\sqrt{p}$). For GB, trees are growing sequentially and each tree is fit on the modified version of the original data set. See Natekin and Knoll[18] for a comprehensive introduction to the GB. In these ways, the trees are less correlated and taking the averages of the results can yield less variance in the prediction. SVM is also adopted given its popularity and flexibility. The current study uses the radial kernel to accommodate the nonlinear boundary between the classes. The prediction of the SVM with the radial kernel is given by:

$$f(x)=\beta_0+\sum_{i\in s}\alpha_i K(x_i,x_{i'}), \tag{8}$$

Where $K(x_i,x_{i'})=\exp\left(-\gamma\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2\right)$.. (9)

The sign of $f(x)$ indicates the classification decision of the test observation x, $\mathcal{S}$ are the support vectors and α_i indicates non-zero weights for all support vectors and otherwise zero. In the radial kernel $K(x_i,x_{i'})$, γ is a positive constant and $\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$ indicates the distance between a training observation and a test observation in terms of the Euclidean distance. Therefore, larger distances lead to smaller radial kernel values; only nearby training observations play a role in the prediction of class labels of a test observation.

Methods

Data

The dataset used in the current study is simulated based on one real dataset provided by the Centers for Disease Control and Prevention[19] as the real data cannot be shared with any party due to confidentiality. The original dataset contains 5,760,066 subjects and 31 variables. For the classification purpose, 25 variables are retained for the analysis. The outcome variable "current_status" indicates the binary status of the subjects (1 = laboratory-confirmed case; 0 = probable cases). The other variables are used as predictors, including race, gender, age group, and survey items, as presented in Table A1 in Appendix A. Complete cases of these 25 variables are retained for the simulation purpose, yielding a sample of 184,567 cases. The proportion of the probable cases is 0.05 in the complete sample, which remains the same as that in the original dataset. This proportion of 0.05 mimics the proportion of positive cases in the realistic COVID diagnosis scenario. Therefore, the prediction goal is to detect the probable cases in the current study in lieu of the positive cases. The simulated dataset was generated as follows. Frist, unique response patterns were obtained from the CDC complete dataset for negative and positive groups, respectively. Second, the new dataset was simulated by sampling the existing response patterns from all survey items proportionally from the CDC complete dataset. Finally, a simulated sample was generated (n = 50,696). The simulated sample was further randomly partitioned into a training sample (70%, n = 35,426) and a test sample (30%, n = 15,270). It is recommended that the size of the training sample be about 2 to 3 times of the size of the test sample to increase the accuracy in prediction.[20]

The development of the classifiers using the above-mentioned machine learning methods consists of two steps: 1) train the model

and determine the tuning parameters using the training data set; 2) evaluate the accuracy of the classifier using the test data set. For the CART technique, the cost-complexity parameter was tuned to find the optimal tree depth using R package *rpart*. The GB method was conducted using R package *gbm*. The tuning parameters included the number of boosting stages, the number of trees, the depth of trees, the learning rate and the minimum number of observations in the tree's terminal nodes. The RF was tuned in terms of the number of predictors sampled for splitting at each node using the R package *randomForest*. A radial basis function kernel SVM was carried out in the current study. Its tuning parameter included the parameter γ in the Equation 9 and the cost value C, which determine the complexity of the decision boundary. After the parameters were tuned, the classifiers were trained using the training dataset. The 10-fold-validation was conducted for CART, GB, RF and SVM in the tuning process to determine the optimal parameter values. Lastly, the trained models were fit on the test dataset to evaluate the classification accuracy. Training and test datasets are scaled for SVM.

Imbalanced classes

An important technical issue in the current study is the extreme imbalanced classes: the dataset has approximately 95% confirmed cases while only 5% probable cases. Such extreme imbalanced classes will affect the prediction accuracy using the classification methods introduced above.[9,21] Therefore, the current study further uses a sampling technique called synthetic minority oversampling technique[21] for balancing the dataset to improve the prediction accuracy. Specifically, the minority class is over-sampled by randomly introducing synthetic examples from the nearest neighbors. In the current study, the minority class (i.e., the probable cases) are over-sampled to match the number of the confirmed cases in the training dataset. This newly generated balanced dataset is used as the new training dataset to train the classifiers. Then, the predictions are made on the original imbalanced validation dataset. It is expected that the test error rates would decrease after the balanced training dataset is used.

Evaluation criteria

The accuracy of the statistical classification methods is evaluated using five outcome measures, namely, overall accuracy, balanced accuracy, sensitivity, specificity, and Kappa. The calculations of these measures are presented as follows:

$$Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives},$$

$$Specificity = \frac{True\ Negatives}{True\ Negatives + False\ Positives},$$

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2},$$

$$Overall\ Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ cases},$$

$$Kappa = \frac{P_o - P_e}{1 - P_e}.$$

Specifically, sensitivity evaluates the performance of the classifier in predicting positive cases; specificity evaluates the performance of the classifier in predicting negative cases; balanced accuracy examines the average accuracy of the classifier in predicting both positive and negative cases; overall accuracy measures the proportion of all correct predictions made by the classifier; and Kappa statistic is a general measure of concordance between predicted and true categories. In its formula, P_o is the observed proportion of agreement, P_e is the proportion of agreement expected by chance. Larger values of these five outcome measures indicate better classification decisions.

Adjusted pooled sampling

The current study proposes a novel pooled sampling method, i.e., the adjusted pooled sampling, based on the prediction results yielded from the statistical classification methods. The specific procedure of the proposed method is described as follows. First, the predicted positive cases will be tested individually based on the assumption that statistical prediction would help identify true positive cases. The predicted negative cases, however, are randomly pooled into small groups and tested as a group. Further, if positive groups are identified, then each subject in that group will be tested individually. The total number of tests needed using this adjusted pooled sampling method will be compared with the traditionally pooled sampling which pools the sample into groups directly without the information of statistical predictions. All the analyses were conducted in the software program RStudio.[22]

Results

The result section is organized as follows. First, the tuning results for the machine learning methods are reported. Second, classification accuracies across methods when extreme imbalanced training dataset was used are presented. Third, classification accuracies when extreme imbalanced training dataset was treated by the SMOTE oversampling procedure are presented. Lastly, the adjusted pooled sampling method based on each statistical method is demonstrated.

The parameter tuning results of the machine learning methods are summarized in Table 1. The models were trained based on these parameters using the original imbalanced training dataset and the oversampled balanced dataset, respectively. Then, the trained models were fit on the same test dataset to obtain the classification accuracy.

Table 1 Model tuning results for machine learning methods

Method	Tuning parameter	Original training dataset (*n* = 35,426)	Balanced training dataset (*n* = 66,990)
CART	Cost-complexity parameter	0.008	0.010
GB	Number of trees	250	75
	Depth of trees	10	10
	Learning rate	0.001	0.001
	Minimum number of observations in terminal nodes	10	5

Table Continued...

Method	Tuning parameter	Original training dataset (n = 35,426)	Balanced training dataset (n = 66,990)
RF	Number of predictors sampled for splitting	7	11
SVM	Cost value C	100	10
	γ	4	4

To calculate the evaluation criteria for the classification accuracy for all methods, a specific cut-off value is necessary. Although 0.5 is a traditionally used cut-off value in classification problems, a more extreme cut-off value may be more appropriate in the current scenario with imbalanced data. The influence of the cut-off values on the performance of all methods using both the imbalanced and oversampled balanced training data set is presented in Figure 1. When the original imbalanced training data was used, for all statistical classification methods, sensitivity (i.e., the correct classification rate of the minority class) dropped dramatically when the cut-off value was larger than 0.05. Overall accuracy and specificity had the opposite trend to the sensitivity. Specifically, overall accuracy and specificity increased when the cut-off value was larger 0.05. This is expected since about 5% of the cases in the training data set are considered as positive cases. When oversampled balanced training dataset was used, sensitivity improved when larger cut-off values were used for all methods except for LCA. This indicates that oversampling approach improved the prediction accuracy of positive cases to some extent. In addition, the oversampling approach yielded better balanced accuracy (i.e., the average of sensitivity and specificity) and Kappa values especially when larger cut-off values are used. The highest Kappa value was less than 0.3, which indicates the overall predictions of all methods were not satisfactory. However, in the scenario of COVID-19 diagnosis, sensitivity is of the most interest because it indicates how well the statistical method can identify positive cases (the minority class). Therefore, a cut-off value of 0.05 was considered as optimal in both the original imbalanced condition and the oversampled balanced condition.

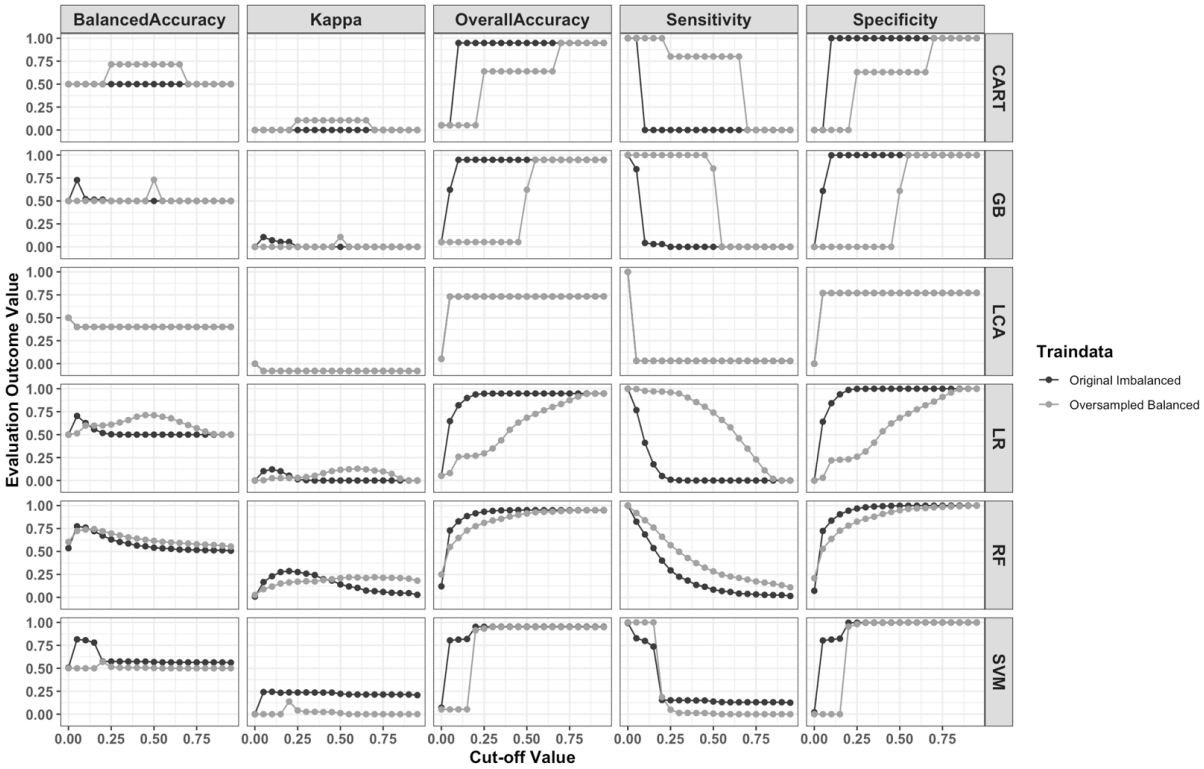

Figure 1 Evaluation criteria under different cut off values for the statistical classification methods.

The prediction results based on the original imbalanced data using the optimal cut-off values (i.e., 0.05) are presented in Table 2. The machine learning methods including CART, RF, GB and SVM achieved better sensitivity compared to traditional logistic regression and latent class analysis. Specifically, CART yielded the highest sensitivity value of 1.000, which indicates that CART has the greatest potential in predicting positive cases in the COVID-19 diagnosis scenario with the extreme imbalanced data structure. In addition, SVM performed the best in terms of balanced accuracy, overall accuracy, specificity and Kappa with a satisfactory sensitivity value of 0.827. The performance of GB, logistic regression, and RF was also acceptable with balanced accuracy, overall accuracy, specificity larger than 0.6. Latent class analysis, however, did not perform well in the extreme imbalanced condition with a small sensitivity value of 0.031.

Table 2 Prediction outcome evaluation of the statistical classification methods

Method	Cut-off Value	Overall Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
LR	0.05	0.647	0.767	0.640	0.704	0.102
LCA	0.05	0.731	0.031	0.770	0.400	-0.079
CART	0.05	0.052	**1.000**	0.000	0.500	0.000
GB	0.05	0.622	0.846	0.610	0.728	0.106
RF	0.05	0.729	0.824	0.723	0.774	0.167
SVM	0.05	**0.804**	0.827	**0.803**	**0.815**	**0.241**

Note: CART; classification and regression tree, GB; gradient boosting, LCA; latent class analysis, LR; logistic regression, RF; random forest, SVM; support vector machine

Largest values in each column are bold face

The prediction results based on the oversampled balanced training data using the optimal cut-off values (i.e., 0.05) are presented in Table 3. A noticeable change is that the sensitivity values of all methods became larger than 0.900 after the training dataset was oversampled to yield a balanced data structure except for LCA. Specifically, CART, GB and SVM all yielded a perfect sensitivity. LCA, however, yielded the same sensitivity of 0.031 as that in the original imbalanced condition. This indicates that the oversampling method improves the performance of statistical classification methods (except for LCA) in terms of detecting the minority group dramatically. As a compromise, specificity values for these methods also dropped compared to that in the original imbalance training data condition.

Table 3 Prediction outcome evaluation of the statistical classification methods after using SMOTE

Method	Cut-off Value	Overall Accuracy	Sensitivity	Specificity	Balanced Accuracy	Kappa
LR	0.05	0.081	0.996	0.031	0.513	0.003
LCA	0.05	0.731	0.031	0.770	0.400	-0.079
CART	0.05	0.052	**1.000**	0.000	0.500	0.000
GB	0.05	0.052	**1.000**	0.000	0.500	0.000
RF	0.05	**0.549**	0.918	**0.528**	**0.723**	**0.089**
SVM	0.05	0.052	**1.000**	0.000	0.500	0.000

Note: CART; classification and regression tree, GB; gradient boosting, LCA; latent class analysis, LR; logistic regression, RF; random forest, SVM; support vector machine

Largest values in each column are bold face

The results of adjusted pooled sampling are shown in Figure 2. Note that the results are based on cut-off values that yielded the smallest number of required tests for all methods. Specific cut-off value information can be found in Table A2 in the Appendix. Six group sizes (5, 10, 15, 20, 50, 100) were manipulated in the current study. In reality, group sizes usually range from 4 to 30 (e.g., Lohse et al., 2020). The current study also investigated relative large group sizes 50 and 100 to further generalize the results. Several observations are made based on Figure 2. First, adjusted pooled sampling method yielded better results than the traditional pooled sampling for all methods except LCA, as shown in Figure 2 that all bars are lower than the dashed lines. Adjusted pooled sampling based on LCA, however, was similar to the traditional methods no matter whether oversampling was used. Second, GB, RF, LR and SVM, yielded better adjusted pooled sampling results given a certain group size. Third, for CART, GB, and LR, adjusted pooled sampling method using oversampled balanced training dataset yielded better results than the method using original imbalanced training dataset, while for SVM, the opposite is true. For RF, adjusted pooled sampling method using oversampled balanced training dataset only yielded better results than the method using original imbalanced training dataset when group size was large (i.e., 100).

Citation: Qiao X, Ding Y, Ying C, et al. Classification and prediction with very imbalanced group sample sizes: an illustration with COVID-19 testing. *Biom Biostat Int J.* 2021;10(4):154–161. DOI: 10.15406/bbij.2021.10.00342

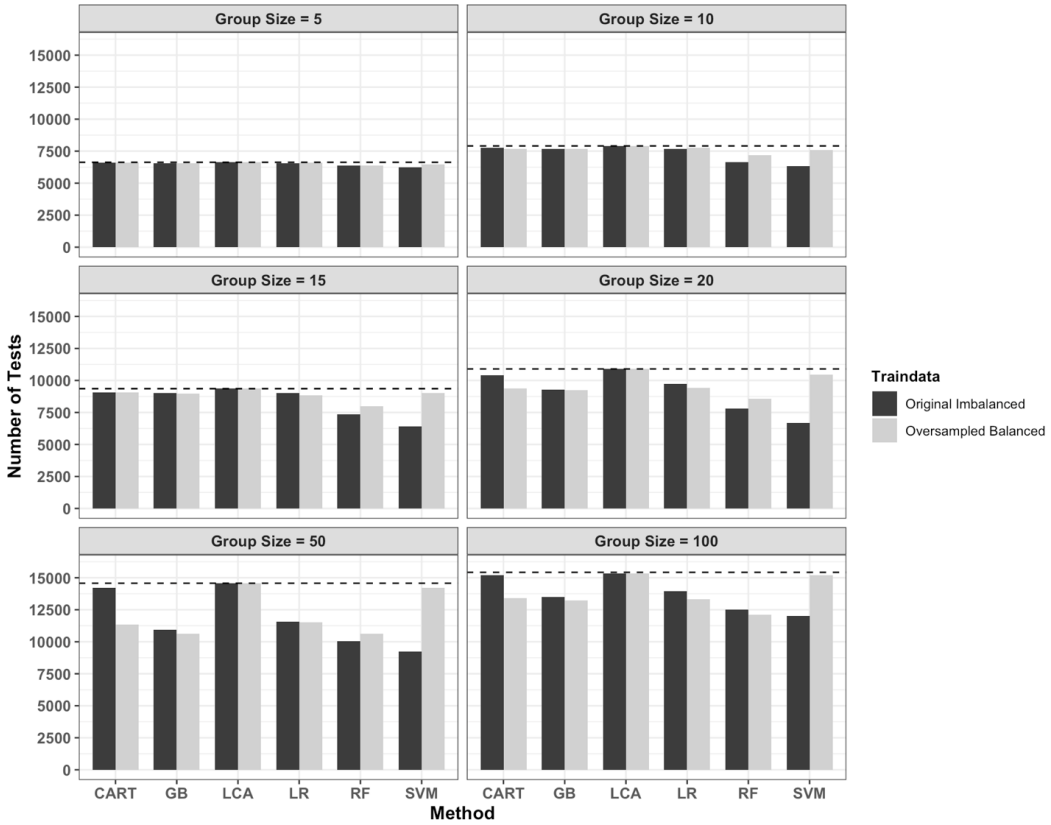

Figure 2 Number of tests required by the adjusted pooled sampling method.

Note: Dashed lines indicate number of tests required by the traditional pooled sampling method.
CART; classification and regression tree, GB; gradient boosting, LCA; latent class analysis, LR; logistic regression, RF; random forest, SVM; support vector machine.

Conclusions and discussion

The current study demonstrates th e applications of statistical classification methods, including logistic regression, latent class analysis, and machine learning methods, on COVID data with extreme imbalance of the outcome variable (confirmed vs. probable cases). The proportion of probable cases is approximately 0.05, which mimics the proportion of confirmed cases in the pandemic. Traditionally, logistic regression is used to conduct classifications on the COVID data.[1] However, latent class models and machine learning methods have not been investigated on COVID data yet. The machine learning methods used in the current study include decision tree, random forest, gradient boosting, and support vector machine. In addition, the current study adopted the oversampling method from the machine learning community to deal with the extreme imbalance scenario. Specifically, a balanced training dataset was generated using the oversampling technique SMOTE to train the machine learning models. Then, the trained models were fit using the same imbalanced test dataset. Furthermore, the current study proposes the adjusted pooled sampling method which utilizes information obtained from the statistical classification methods to facilitate the COVID testing efficiency.

To illustrate the statistical classification methods, a simulated dataset based on the real dataset obtained from the Centers for Disease Control and Prevention was used. The results indicated that 1) machine learning methods outperformed the logistic regression model and latent class analysis in terms of sensitivity, i.e., the accuracy of detecting the minority group, under the extreme imbalance data scenario; 2) the choice of cut-off value is related to the proportion of the two groups in the outcome variable. For example, in the extreme imbalance scenario, a cut-off value of 0.05 yielded the highest sensitivity for logistic regression, CART, GB, RF and SVM rather than the traditional cut-off value 0.5; 3) after the extreme imbalance was treated in the training dataset, the sensitivity increased for LR, CART, RF, and SVM; 4) adjusted pooled sampling method based on prediction results from the statistical methods (e.g., SVM) requires less number of COVID tests than traditional pooled sampling method.

The current study is a demonstration of the statistical classification methods on COVID-19 data. Despite the promising results, several limitations exist in the current study. First, the dataset used in the current study includes the outcome variable of confirmed cases vs. probable cases rather than positive cases vs. negative cases in the real scenario. However, the proportions of the confirmed cases and probable cases in the dataset are similar to those of positive cases and negative cases in real COVID-19 diagnostic settings. Second, the adjusted pooled sampling share the same limitation with the regular pooled sampling. Although pool sizes larger than 10 yielded better testing efficiency, large pool sizes may dilute the specimen and lead to higher false negative rates. Future studies are needed to examine whether the proposed method has better performance when the prevalence is low and larger pools can be formed. Lastly, the performance of the statistical classification methods demonstrated in the current study is based on the COVID context. However, the method can be generalized to any scenario where extreme imbalance of the dataset exists.

Citation: Qiao X, Ding Y, Ying C, et al. Classification and prediction with very imbalanced group sample sizes: an illustration with COVID-19 testing. *Biom Biostat Int J.* 2021;10(4):154–161. DOI: 10.15406/bbij.2021.10.00342

Acknowledgments

None.

Conflicts of interest

None.

References

1. Bhargava A, Fukushima EA, Levine M, et al. Predictors for severe COVID–19 infection. *Clin Infect Dis*. 2020;71(8):1962–1968.
2. Jehi L, Ji X, Milinovich A, et al. Individualizing risk prediction for positive coronavirus disease 2019 testing: results from 11, 672 patients. *Chest.* 2020;158(4):1364–1375.
3. Marin EC, Buld L, Theiss M. Connectomics analysis reveals first, second, and third–order thermosensory and hygrosensory neurons in the adult drosophila brain. *Curr Biol.* 30:3167–3182.
4. Wynants L, Van Calster B, Collins G S, et al. Prediction models for diagnosis and prognosis of covid–19: Systematic review and critical appraisal. *BMJ*. 2020;369:m1328.
5. Yinxiaohe Sun, Vanessa Koh, Kalisvar Marimuthu, et al. National Centre for Infectious Diseases COVID–19 Outbreak Research Team, Epidemiological and Clinical Predictors of COVID–19, *Clin Infect Dis*. 2020;71(15):786–792.
6. Akaike H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 1974;19(6):716–723.
7. Kumar R, Arora R, Bansal V, et al. Accurate prediction of COVID–19 using chest X–Ray images through deep feature learning model with SMOTE and machine learning classifiers. *MedRxiv.* 2020.
8. Hassanien A E, Mahdy L N, Ezzat K A, et al. Automatic x–ray covid–19 lung image classification system based on multi–level thresholding and support vector machine. *MedRxiv*. 2020.
9. King G, Zeng L. Logistic regression in rare events data. *Political Analysis.* 2001;9(2):137–163.
10. EI–Elimat T, AbuAlSamen MM, Almomani BA. Acceptance and attitudes toward COVID–19 vaccines: A cross–sectional study from Jordan. PLoS ONE 16(4):e0250555.
11. Lazarsfeld PF, Henry NW. Latent structure analysis. *Houghton, Mifflin*. 1968.
12. Everitt BS. An introduction to finite mixture distributions. *Stat Methods Med Res* . 1996;5(2):107–127.
13. Dayton C, Macready G. Concomitant–Variable Latent–Class Models. *Journal of the American Statistical Association*. 1988;83(401):173–178.
14. James G, Witten D, Hastie T, et al. An Introduction to Statistical Learning, Vol 112. New York, NY: *Springer*. 2013.
15. Qiao X, Jiao H. Data mining techniques in analyzing process data: a didactic. *Front Psychol.* 2018;9:2231.
16. Breiman L. Some properties of splitting criteria. *Machine Learning*. 1996;24(1):41–47.
17. Kuhn M. Predictive Modeling with R and the caret Package. 2013.
18. Natekin A, Knoll A. Gradient boosting machines, a tutorial. *Front Neurorob*. 2013;7:21.
19. Centers for Disease Control and Prevention COVID–19 Response. COVID–19 Case Surveillance Data Access, Summary, and Limitations (version date: June 27, 2020). 2020.
20. Sinharay S. An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*. 2016;35:38–54.
21. Chawla N V. Data mining for imbalanced datasets: An overview. *Data mining and knowledge discovery handbook*. 2009;875–886.
22. R Studio Team. R Studio: Integrated development environment for R. 2017.