International Journal of
Statistics and Applied Mathematics

ISSN: 2456-1452
Maths 2023; 8(5): 58-62
© 2023 Stats & Maths
https://www.mathsjournal.com
Received: 20-06-2023
Accepted: 23-07-2023

Preethi Jayarama Shetty
Lecturer, Department of PG Studies and Research in Statistics, Mangalore University, Mangalore, Karnataka, India

Satyanarayana
Lecturer, Department of PG Studies and Research in Statistics, Mangalore University, Mangalore, Karnataka, India

Corresponding Author:
Preethi Jayarama Shetty
Lecturer, Department of PG Studies and Research in Statistics, Mangalore University, Mangalore, Karnataka, India

Prediction performance of classification models for imbalanced liver disease data

Preethi Jayarama Shetty and Satyanarayana

Abstract
The liver is the largest solid organ that plays an important role in many bodily functions from protein production and blood clotting to cholesterol. It additionally serves to eliminate harmful waste products and certain drugs, detoxify alcohol, and environmental toxins. The liver forms and secretes digestive fluid that contains digestive fluid acids to help with the digestion, internal organ absorption of fats, and fat-soluble vitamins A, D, E, and K. Diseases that may affect the liver include hepatitis, cirrhosis, fatty liver, and liver cancer. The processing of unbalanced data presents significant difficulties for the researchers in class label identification. In health research, accurate disease diagnosis employing a good classification model would alleviate the strain on doctors and help to prevent major losses. The main focus of the study is to identify the major risk factors associated with liver disease and identifying the best classification model to handle imbalanced liver disease data. The prediction performance of different classification models using SMOTE and ROSS algorithm are compared based on accuracy measures and the best classification model to deal with imbalanced data is reported.

Keywords: Imbalanced data, classification, accuracy, SMOTE, ROSS, Prediction

1. Introduction

The liver regulates most chemical levels within the blood and excretes a product called bile. This helps take away waste products from the liver. All the blood effort the abdomen and intestines pass through the liver. the disease will be genetic and Liver issues may also be caused by a variety of factors that injury the liver, such as viruses, alcohol use, and fatness. Over time, conditions that injury the liver will result in scarring (cirrhosis), which might result in liver failure, a dangerous condition. however early treatment could offer the liver time to heal.

Hepatitis: Hepatitis is inflammation of the liver and is sometimes caused by viruses like hepatitis A, B, and C. This illness will
have non-infectious causes like serious drinking, drugs, allergies, or obesity.

Cirrhosis: Long-term damage to the liver from a variety of causes can leading to permanent scarring and liver failure, Known as cirrhosis.

Liver cancer: Hepatocellular carcinoma is the most common type of liver cancer and occurs after cirrhosis is present.

Liver failure: Infection, genetic diseases, and excessive alcohol can lead to liver failure.
Ascites: cancer cells irritate the lining of the tummy, the liver leaks fluid into the belly, which becomes distended and heavy.

- **Gallstones:** If a gallstone becomes stuck in the bile duct draining the liver, hepatitis and bile duct infection (Cholangitis) can result.
- **Hemochromatosis:** Hemochromatosis permits iron to deposit within the liver, damaging it. The iron conjointly deposits throughout the body, inflicting multiple alternative health issues.
- **Primary sclerosing cholangitis:** A rare sickness with unknown causes, primary sclerosing cholangitis causes inflammation and scarring within the gall ducts within the liver.
- **Primary biliary cirrhosis:** During this rare disorder, the associate unclear method slowly destroys the gall ducts within the liver. It may develop eventually.
- This study aims to detect the presence of a disease using a good classification model, which will reduce the burden of doctors as well as avoid great loss.

2. Literature review

In real life situation most of health-related data are highly imbalanced and ignorance of imbalanced situation while constructing classifier may leads to wrong conclusion. Adnan Amin (2016) [2] studied the performance of different oversampling techniques with evaluation of four rules-generation algorithms and the empirical results demonstrate that the overall predictive performance of MTDF and rules generation based on genetic algorithms performed the better as compared with the rest of the evaluated oversampling methods and rule-generation algorithms. In health science, correct detection of presence of a disease using good classification model will reduce burden of doctors as well as avoid great loss. Comparative Study on Liver Disease Prediction Using Supervised Machine Learning Algorithms was carried out by A.K.M Sazzadur Rahman *et al*. (2019) [1]. Here he compares the Logistic Regression with machine learning classifiers for predicting chronic liver disease and obtain Logistic Regression classification technique is more effective.

3. Materials and Methods

In real life, most of the categorical dataset are imbalance. Handling of imbalanced data possesses great challenges for the researchers to identify the class labels. In health science, correct detection of presence of a disease using good classification model will reduce burden of doctors as well as avoid great loss. In this study, we handle imbalanced liver disease data using SMOTE and ROSS sampling techniques.

3.1 ROSS: Random over-sampling is oversampling technique used to balance the imbalanced dataset. In ROSS, new minority samples are created by randomly selecting training samples from minority class, and then duplicating it. In doing therefore, the category distribution is often balanced, however this could typically cause over-fitting and longer training time throughout imbalance learning method.

3.2 SMOTE: To overcome the issue of over-fitting and extend the decision area of the minority class samples, a novel technique SMOTE ‘‘Synthetic Minority Oversampling Technique’’ was introduced by Chawla, this technique produces artificial samples by using the feature space rather than data space. It is used for oversampling of minority class by creating the artificial data instead of using replacement or randomized sampling techniques. It was the first technique which introduced new samples in the learning dataset to enhance the data space and counter the scarcity in the distribution of samples. The oversampling technique is a standard procedure in the classification of imbalance data (e.g., minority class).

4. Classification models

4.1 Logistic regression

Multivariable methods of statistical analysis appear in general health science literature. In the strict sense, multivariate analysis refers to simultaneously predicting multiple outcomes and it’s uses multiple variables to predict a single outcome. The multivariable methods are a relation between two or more regressors and one responding variable. The model identifies the relationship between dependent variable and several independent variables with corresponding coefficient of the regressors along with error term. The coefficients are the best mathematical fit for the specified model. A coefficient indicates the impact of each regressor on the dependent variable adjusting for all other independent variables. The model serves two purposes: (1) it can predict the value of the responding variable for new values of the regressors, and (2) it can help describe the relative contribution of each regressor to the responding variable, dominant for the influences of the other regressors. The logistic regression is the most popular multivariable method employed in health science.

4.1.1 Fitting the logistic regression model

Although logistic regression model,

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X \qquad (4.1.1.1)$$

looks similar to a simple linear regression model, the underlying distribution is binomial and the parameters, α and β cannot be estimated using least square method. Using the method of maximum likelihood of observing the sample values, the parameters are estimated. A “likelihood” is a probability that the observed values of the response variable can be predicted from the observed values of the control variables. In logistic regression, we observe categorical outcome and independent variables, and that we would like to draw inferences concerning the probability of an event in the population. In a population from which we are sampling, for each individual in our sample of size n, Yi =1 indicates that an event occurs for the i^{th} subject with probability ‘p’, otherwise,

Yi =0. The observed data are $Y_1, \dots \dots, Y_n$ and $X_1, \dots \dots, X_n$

The joint probability of the data is given by

$$L = \prod_{i=1}^{n} \pi(x)^{Yi} (1 - \pi(x))^{1-Yi} (4.1.1.2)$$

Natural logarithm of the likelihood is

$$l = log(L) = \sum_{i=1}^{n} Y_i \, log\big(\pi(x)\big) + (n - \sum_{i=1}^{n} Y_i) log\big(1 - (x)\big) \qquad (4.1.1.3)$$

In which

$$\pi(x)=p(y/x)=\frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \qquad (4.1.1.4)$$

To find the parameters that maximises L. We differentiate L w.r.t α & β and equate into zero.

$$\sum_{i=1}^{n}\left(y_i - \pi(x)\right) = 0 \qquad (4.1.1.5)$$

$$\sum_{i=1}^{n} x_i\left(y_i - \pi(x)\right) = 0 \qquad (4.1.1.6)$$

In Logistic regression equations (5) & (6) are non- linear. Thus we use special method for their solution these methods are iterative in nature. In similar way we estimated parameters $\beta_1, \beta_2, \dots \dots \dots, \beta_k$ in multivariate logistic regression.

4.1.2 Evaluation of a logistic regression model

In this, initially the general model should be assessed and the importance of each of the regressor should be assessed. Then, accuracy or discriminating ability of the model should be evaluated. Finally, the model should be valid.

Overall model evaluation

Using likelihood ratio test the overall fit of the model with k coefficients can be examined. The null hypothesis is,

$$H_0 = \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

The difference of these two yields a goodness of fit index G, χ2 statistic with k degrees of freedom. This is a measure the effect of independent variables on dependent variable.

$$G = \chi 2 = -2log\frac{likelihood\ of\ the\ null\ model}{likelihood\ of\ the\ given\ model} \qquad (4.1.2.1)$$

If the p value for the overall model fit statistic is less than 0.05, then reject $H_0$and conclude that there is evidence that at least one of the independent variables contributes to the prediction of the outcome.

Statistical significance of individual regression coefficients: The logistic regression coefficient for the ith independent variable is changed1 unit while all of the other predictors are held constant, log odds of outcome is expected to change by units. The likelihood ratio test and the Wald statistic are designed to assess the significance of an independent variable in logistic regression.

Wald statistic: The Wald statistic can be used to find out the significance of individual coefficients in a given model. The Wald statistic is

$$W_j = \frac{\hat{\beta}_j^2}{SE\left(\hat{\beta}_j^2\right)} \qquad (4.1.2.2)$$

Each W_j is compared with $\chi^2_{(1)}$.

4.1.3 Receiver Operating Characteristic (ROC) Curve

ROC curve abbreviates the model's performance by evaluating the connection between sensitivity and 1-specificity, as the value of the cut-point cc is increased from 0to 1. High sensitivity and specificity may lead to a model with high discrimination ability and it implies an ROC curve goes close to the top left corner of the plot. The ROC curve has 45-degree diagonal line means a model have no discrimination ability. The area under curve (AUC), referred to as index of accuracy, is a perfect performance metric for ROC curve. High AUC leads to better the prediction power of the model.

4.2 Decision tree

A decision tree could be a flowchart-like tree structure, where the uppermost node in tree is the root node, test on an attribute denotes by every internal node, an outcome of the test represented by every branch, and each terminal node holds a class label. Given a tuple which associated unknown class label, the attribute values of the tuple are tested against the choice tree. To find the class expected values for that tuple, a path is traced from the root to a terminal node. Since the construction of the decision tree requires neither domain expertise nor parameter setup, it is excellent for exploratory knowledge discovery. Data with several dimensions can be handled using decision trees. The predicted value for each of those tuples is regarded to be the average of the values of the dependent variables in each tuple. The attribute selection criterion used by the CART algorithm is the Gini index.

4.3 Random forest

Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. To correct over fitting problem of decision tree, Random decision forest is the alternative. Random forest adds additional randomness to the model, while growing the trees. Random forest searches for the best feature among a random subset of features instead of searching for the most important feature while splitting a node. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. Another great quality of the random forest algorithm is that it is very easy to measure the importance of each feature on the prediction. By looking at the feature importance you can decide which features to possibly drop because they don't contribute enough to the prediction process. The hyper parameter in random forest are used to increase the predictive power of the model as well as the speed of model.

4.4 Support vector machine

It can be used to solve classification and regression issues. By generating a hyperplane, this technique divides the data into two classes. SVM is an algorithm that accepts data as input and outputs, if possible, a line that divides those classes. The data points from both classes that are closest to the line are called as support vectors by the SVM method. Then, compute the distance between the line and the support vectors, this distance is called as margin. The goal is to maximise the profit margin. The greatest margin merely refers to the optimal hyperplane. As a result, SVM attempts to generate a decision boundary with the greatest feasible difference between the two groups.

4.5 Naive bayes classifier

Bayesian classifiers are statistical classifiers. This is the algorithm to predict the probability of given tuple that belongs to a particular class. The Bayes theorem is the foundation of Bayesian classification. A simple Bayesian classifier is known as the 'naïve Bayesian classifier' to be comparable in performance with decision trees and selected neural network classifiers. When naïve Bayesian classifier applied to a large database, it gives high accuracy and speed. According to this algorithm, effect of an predictor on a given class is purely independent of the values of the other attributes. It is known as class conditional independence. To predict the class label of X, $P(X|C_i)P(C_i)$ is evaluated for each class Ci. The classifier predicts that the class label of X is Ci if and only if it is the class that maximizes $P(X|C_i)P(C_i)$.

4.6 K-NN Classification

KNN compare the given test tuple with training tuple which is similar to it, so we can say it is based on learning by analogy.

Once given an unknown tuple, a K- nearest neighbour classifier search the pattern area for the k-training tuple that area unit nearest to the unknown tuple. Closeness is outlined in terms of distance matric. For K-nearest neighbour classification, the unknown tuple is allotted to the foremost category among its K-nearest neighbours. K-nearest neighbour classifiers can also be used for prediction, that is, to return a real valued prediction for a given unknown tuple. A good value for K, number of nearest neighbours, can be found experimentally or k may be taken as

$$k = \sqrt{number\ of\ training\ tuples}$$

4.7 Classification table

The confusion matrix is a method to evaluate accuracy of the logistic regression.

Predicted class Actual class	**Class 1**	**Class 2**
Class 1	True Positive (TP)	False Negative (FN)
Class 2	False Positive (FP)	True Negative (TN)

Evaluation measures

- **Accuracy:** It is a measure that calculates the classifier's overall accuracy.

It is formulated as:

$$\boldsymbol{Accuracy} = \frac{TN\ +\ TP}{\mathrm{TN + FN + FP + TN}}$$

- **Sensitivity (Recall):** It is the proportion of those cases which are correctly classified as true positive, and calculated as:

$$Recall\ = \frac{TP}{FN + TP}$$

- **Specificity:** It is the proportion of those cases which are correctly classified as true negative, and calculated as:

$$Specificity = \frac{TN}{FP + TN}$$

- **Precision:** It is number of positive class predictions that actually belong to the positive class. Formally, it can be expressed as;

$$Precision\ = \frac{TP}{\mathrm{TP\ +\ FP}}$$

- **F-Measure:** It is based on the weighted harmonic mean between both the precision and recall. If both precision and recall are reasonably high that gives high F-measure value. It can also be considered as the weighted-average of recall and precision.

$$F - Measure\ = \frac{2\ \times\ Recall\ \times\ Precision}{\mathrm{Recall\ +\ Precision}}$$

5. Data analysis and results

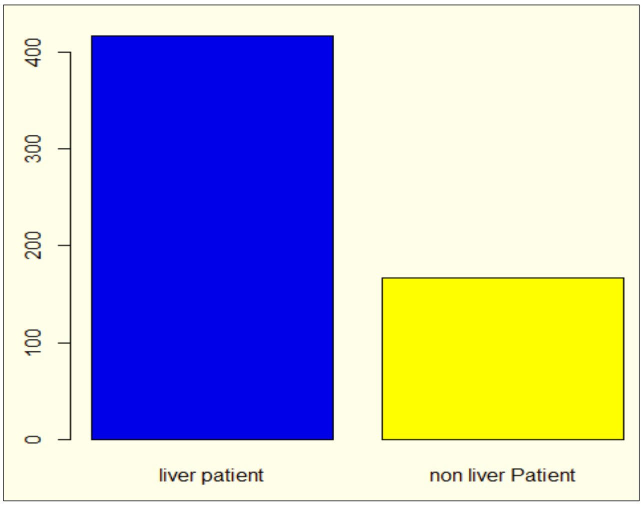

Fig 1: Plot of ratio of liver patients

In real life most of the categorical data are imbalanced. This **liver disease** data contains following variables, Age, Gender, Total Bilirubin, Direct Bilirubin, Alkaline Phosphotase, Alamine Aminotransferase Aspartate Aminotransferase, Total proteins, Albumin, Albuminand Globulin Ratio and class. From the data, we observe that there were 583 liver patients and it contain 441 male patient records and 142 female patient records in which 416 were having liver disease. Count Plot of ratio of liver patients is shown in figure

Table 1: shows the Model parameters, Significant regressors and its coefficient values.

	Estimate	**Std. Error**	**Z value**	**Pr(>|z|)**
(Intercept)	3.5587256	1.2980427	2.742	0.006114
Age	-0.0187478	0.0063646	-2.946	0.003223
Direct_Bilirubin	-0.4925101	0.1730652	-2.846	0.00443
Alkaline_Phosphotase	-0.0013123	0.0008125	-1.615	0.006259
Alamine_Aminotransferase	-0.0147287	0.004001	-3.681	0.000232
Total_Protiens	-0.9267711	0.3715429	-2.494	0.012618
Albumin	1.7158457	0.7259655	2.364	0.018101
Albumin_and_Globulin_Ratio	-1.8340476	1.1052096	-1.659	0.047024

Table 2: Performance evaluation of different classification models under different sampling techniques

Method		Accuracy	Sensitivity	Specificity	Precision	F-measure
Random sampling	Logistic model	0.70	0.85	0.34	0.76	0.80
	Decision tree	0.65	0.84	0.16	0.71	0.77
	Naive bayes	0.56	0.44	0.86	0.89	0.59
	K-NN C	0.64	0.76	0.34	0.74	0.75
	Random forest	0.68	0.83	0.30	0.75	0.79
	SVM	0.71	0.97	0.08	0.72	0.84
ROSS oversampling	Logistic model	0.60	0.54	0.76	0.85	0.66
	Decision tree	0.59	0.62	0.52	0.76	0.69
	Naive bayes	0.54	0.38	0.92	0.92	0.54
	K-NN	0.55	0.50	0.68	0.80	0.62
	Random forest	0.70	0.82	0.40	0.77	0.80
	SVM	0.60	0.53	0.78	0.86	0.63
SMOTE oversampling	Logistic model	0.61	0.58	0.68	0.82	0.68
	Decision tree	0.63	0.70	0.48	0.77	0.73
	Naive bayes	0.57	0.46	0.86	0.89	0.60
	K-NN	0.55	0.54	0.56	0.76	0.64
	Random forest	0.62	0.67	0.48	0.77	0.74
	SVM	0.65	0.60	0.76	0.86	0.68

According to Table 2, SVM under SMOTE oversampling performs better compare to all other classifiers based on all evaluation measures

6. Conclusion

In real life, most of the categorical dataset are imbalance. Imbalanced data possesses great challenges for the researchers to identify the class labels. Most of the classification models based on random sampling yields either low sensitivity or specificity, because it is biased towards majority class. In health science, correct detection of presence of a disease using good classification model will reduce burden of doctors as well as avoid great loss. In this study, we handle imbalanced liver disease data using SMOTE and ROSS sampling techniques. In ROSS and SMOTE sampling, class distribution can be balanced using duplicate class and artificial class respectively. The result shows that prediction performance of all the seven classifiers has been improved by considering ROSS and SMOTE sampling techniques. Based on overall performance, SVM under SMOTE over sampling is the best classifier to deal with imbalanced liver disease data.

7. Reference

1. Sazzadur Rahman AKM, Javed FM, Shamrat M, Tasnim Z, Roy Z, Hossain SA, *et al*. Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms. International journal of scientific & technology research. 2019, 8(11).
2. Amin A, Anwar S, Adnan A, Nawaz M, Howard N, Qadir J, *et al*., Comparing Oversampling Techniques to Handle the Class Imbalance Problem: A Customer Churn Prediction Case Study. IEEE Access; c2016.
3. Andrew Wong, Mohamed S. Kamel, Classification of imbalanced data. International Journal of Pattern Recognition and Artificial Intelligence; c2011.
4. David W. Hosmer and Stanley Lemshow: Applied Logistic Regression, Second Edition, Wiley Series in Probability and Statistics; c2000.
5. Han, Kamber, Pei. Data Mining Concepts and Techniques, Third Published by Elsevier; c2013.