Bootstrapped Multinomial Logistic Regression on
Apnea Detection Using ECG Data

Hadaiq R Sanabila, Mohamad Ivan Fanany, Wisnu Jatmiko, and Aniati Murni Arimurthy
Laboratory of Pattern Recognition, Image Processing, and Content Based Image Retrieval
Faculty of Computer Science, Universitas Indonesia
Email: hadaiq.rolis@ui.ac.id, ivan@cs.ui.ac.id, wisnuj@cs.ui.ac.id, aniati@cs.ui.ac.id

Abstract—In designing a classification system, one of the most important considerations is how optimal the classifier will adapt and give best generalization when it is given data from unknown model distribution. Unlike linear regression, logistic regression has no simple formula to assess its generalization ability. In such cases, bootstrapping offers an advantage over analytical methods thanks to its simplicity. This paper presents an analysis of bootstrapped multinomial logistic regression applied on apnea detection using ECG data. We examine multinomial logistic regression to detect or recognize multi-class apnea categories (heavy, middle, healthy). We show that for generally complex and highly unstructured medical data such as ECG for apnea detection, bootstrapping gives more meaningful assessment to detect over-fitting than k-fold cross validation. The bootstrapping also gives higher classification accuracy prediction over k-fold cross validation for the same training data proportion.

I. INTRODUCTION

Nowadays statistical methods has become a very powerful tool for supporting medical decisions. Doctors can be helped by statistical models to interpret correctly so many data and to support their decisions. Even further, due to the characteristic of medical data and huge number of variables to be considered, researches to automate the decision support system that can be performed by machine (e.g., detection apnea symptoms using computers at home [1], or wearable alarm clock that can detect sleep stages [2]) is continuously proliferating. Of course, although the statistical models are very powerful for doctors and computers are getting faster, the models and the computers cannot substitute doctors’ viewpoint.

Many of the medical problems are related to questions of classification and prediction, many times with only two categories (disease or not disease, for instance). In those cases, the use of classical techniques has to be restricted to some specific methods such as logistic regression or similar. When we are dealing with multi-categories classifications usually we prefer to use newer tools such as neural networks which do not require such restrictive hypothesis. In addition, our preference to such newer tools is also due the fact that medical data are generally complex and have lack of structure. However, according to a study [3] which compared logistic regression and neural networks, several points should be remarked:

- For data that was used to build the model, some of the neural networks can be considered better, but for data that was used to validate the model none of them are better.
- The neural networks models that can be considered better in building data, actually needs a greater number of inputs.
- Doctors usually prefer to use logistic regression which includes less variables and have an easy meaning in the model.
- Prediction can be done with a simple equation in logistic regression, whereas the necessity of computing tool in neural network can delay the diagnose for the doctors.

Hence, [3] recommends to use logistic regression for medical data. Recently, [4] also use multinomial logistic regression to examine the influence of age, body mass index (BMI), gender, smoking, and social characteristics on apnea hypopnea index (AHI).

Cross validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimates how accurately a predictive model will perform in practice. One round of cross validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset over the rounds.

In this paper, we try to asses and estimate how accurately the multinomial logistic regression classifier will perform in practice in terms of classification accuracy and generalization ability by analyzing classification of ECG data samples to automatically detect the disordered breathing during sleep (apnea). We will show that bootstrapping cross validation method is preferred than k-fold cross validation for the evaluation and assessment of this data samples.

II. SLEEP APNEA

Sleep disorder is a medical disorder in sleep of a
person or animal. It can be classified into dyssomnias (characterized by either hypersomnolence or insomnia) and parasomnias (characterized by abnormal and unnatural movements in sleep). Sleep disorder interferes with normal physical, mental and emotional functioning. One category of dyssomnias is sleep apnea. The sleep apnea is defined as abnormal pause of airflow during sleep, preventing air from entering the lungs and causes an obstruction. The types of sleep apnea are central sleep apnea (CSA), obstructive sleep apnea (OSA), and mixed sleep apnea (both central sleep apnea and obstructive sleep apnea). CSA occurs when the brain does not send the order signal to the muscles to take a breath so there is no muscular effort to take a breath. OSA happens when the brain sends the instruction signal to the muscles and the muscles make an effort to take a breath but they are unsuccessful because the airway becomes obstructed and prevents an adequate flow of air. Mixed sleep apnea, occurs when both central sleep apnea and obstructive sleep apnea are occurred.

Sleep disorder is related to a risen risk of cardiovascular illness, stroke, high blood pressure, arrhythmias, diabetes, and sleep deprived driving accidents [1-4]. Sleep apnea person also have a 30% higher risk of heart attack or premature death than those unaffected [5].

The symptoms of sleep apnea may include daytime fatigue & sleepiness, insomnia, poor concentration, memory problems, anxiety, irritability, headaches, and impediment performing work duties. Generally, sleep disorder is diagnosed using polysomnoogram. It records minimum up to twelve channels: three channels for EEG (Electroencephalography), one or two channels for quantify airflow, one channel for chin muscle tone, one or more channels for leg movement, two for eye motion, one for heart rate, one for oxygen intensity, and one for each belts which measures chest wall movement and upper abdominal wall movement. The polysomnographic diagnose method is relatively costly and due to the dearth of diagnostic sleep laboratories, the sleep apnea is widely under-diagnosed [6]. In addition, polysomnoogram does not provide a sound prove that sleep apnea as indirect of evidence of airflow measurement and respiratory movement that is used to evaluate the disordered breathing. Therefore we want to provide a reliable identification of disordered breathing with fewer and simpler measurement using ECG.

III. LOGISTIC REGRESSION

A. Logistic Regression

Logistic regression is one of the regression methods which is used to predict the probability of dichotomous event. It is used when the dependent variable is a dichotomy and the independent variables are of any type. Usually, logistic regression is used to predict a dependent variable on the basis of continuous and categorical independents. It is also used to determine the effect of the size of the independent variables on the dependent. Odds ratios are used to clarify the impact of predictor variables.

Logistic regression use logistic function as a model. Logistic function is useful because it can take as an input any value from negative infinity to positive infinity, whereas the output is confined to values between 0 and 1. The formula of logistic function is described below:

$$f(z) = \frac{e^z}{1 + e^z}$$

The variable $z$ defines as $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$, where $\beta_0, \beta_1, \ldots, \beta_k$ called the intercept and the $x_1, x_2, \ldots, x_k$ called as coefficient. The coefficient describes the proportions of that risk factor, a positive coefficient means that the variable increase the probability of outcome and vice versa.

$$\text{odds} = \frac{e^z}{1 + e^z}$$

where $z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$.

Logistic regression is a suitable way of describing the relationship between independent variables and a binary response variable, expressed as a probability that has dichotomous events.

B. Multinomial Logistic Regression

Multinomial logistic regression is the generalization of logistic regression which allowing more than two outcomes events. Multinomial logistic regression used when the dependent variables cannot be ordered in any meaningful way. Its model assumes that data in each independent variable has a single value for each case.

Multinomial logistic regression selects one of dependent variable as the confrontation category and assumes that the dependent variable cannot be perfectly predicted from the independent variables. Multinomial logistic regression compares multiple groups through a combination of binary logistic regressions. The models of multinomial logistic regression are

$$P_{y_i = j} = \frac{e^{\beta_j x_i}}{1 + \sum_{k=1}^{K} e^{\beta_k x_i}}$$

and

$$P_{y_i = j} = \frac{e^{\beta_j x_i}}{1 + \sum_{k=1}^{K} e^{\beta_k x_i}}$$

where for the $i^{th}$ individual, $y_i$ is the outcome event and $X_i$ is a vector of explanatory/independent variables. The parameters $\beta_j$ are typically approximated by maximum a posteriori (MAP) which is an extension of maximum likelihood using regularization of the weight to prevent morbid.
solutions. The solution is typically found using and iterative procedure such as iteratively re-weighted least squares (IRLS) or, more commonly these days, a quasi-Newton method such as the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) method.

Multinomial logistic regression does not make any assumption of normality, linearity, and homogeneity of variance for the independent variables. Because it does not impose these requirements, it is preferred to discriminant analysis when the data does not satisfy these assumptions.

IV. CROSS VALIDATION

Cross validation is important in guarding against testing hypotheses suggested by the data, especially where further samples are hazardous, costly or impossible to collect. In linear regression where the training mean-squared-error (MSE) underestimates the validation MSE, cross validation is not practically useful. In most other regression procedures like logistic regression, however, no simple formula to make such adjustment. Cross validation is a generally applicable way to predict the performance of a model on a validation set using computation in place of mathematical analysis. There are several types of cross validation: (1) K-fold cross validation, (2) Leave-one-out or Jack-knife cross validation, (3) Repeated random sub-sampling or bootstrapping. In this study, we compare bootstrapping to K-fold cross validation.

A. K-fold

K-fold cross validation is one cross validation method which divide dataset into k datasets and the holdout method is iterated k times. The cross-validation K-Fold is the data resulted from the observation of the mentioned sub-groups to be K which denotes the number of sub-groups. A sub-group of data is used for validation and testing and the rest sub-group K – 1 is for training process. The mentioned process is repeated for K times with each sub-group used one time only. The result of such calculation is the mean value to get the final value. The advantage of this method is that all observed data is used either in training or testing process. The conducted experiment uses 3, 5, 10, 15, and 20 folds cross validation.

B. Repeated Random Sub-sampling (Bootstrapping)

Bootstrapping is one of the resampling methods used to measure the properties of estimator from the distribution approximation sample. This method estimates the error of the median by raffling the sample with replacement from original data and estimate the standard error of the original median from the observed variable. Bootstrapping is a reliable alternative to deduct based on parametric assumptions when those assumptions are in doubt or parametric deduction is requires very complicated formulas for the standard error calculation.

In [17], Ader et al. (2008) suggests to employ bootstrapping for the following condition:

- When the theoretical distribution of a statistic is complicated or unknown
- When the sample size is insufficient for straightforward statistical inference
- When power calculations have to be performed, and a small pilot sample is available

V. EXPERIMENTAL SETUP

The database used in this experiment was obtained from PhysioNet [6]. It is recorded from 35 persons containing single ECG signal of approximately 8 hours observation which divided into one-minute segments. The ECG signals were extracted from recorded polysonogram measurement. Each segment annotated based on 3 classifications that is Apnea (A), Borderline/mild apnea (B), and Normal (N). The total number of segments classified in this manner was 16907 with 10415 (61.6%) annotated as normal, 252 (1.5%) annotated as mild apnea and 6240 (36.9%) annotated as apnea.

The features used in this experiment was extracted from ECG signal using QRS detector program [7][8]. The considered features were:

- Mean of RR-interval,
- Standard deviation RR-interval,
- The NN50 measure (variant 1), defined as the number of pairs of adjacent RR-intervals where the first RR-interval exceeds the second RR-interval by more than 50 ms,
- NN50 measure (variant 2), defined as the number of pairs of adjacent RR-intervals where the second RR-interval exceeds the first RR-interval by more than 50 ms,
- Two pNN50 measures, defined as each NN50 measure divided by the total number of RR-intervals,
- the SDSD measures, defined as the standard deviation of the differences between adjacent RR-intervals,
- RMSSD measure, defined as the square root of the mean of the sum of the squares of differences between adjacent RR-intervals,
- Median of RR-interval,
- Inter-quartile range, defined as difference between 75th and 25th percentiles of the RR-interval value distribution,
- Mean absolute deviation values, defined as mean of absolute values obtained by the subtraction of the mean RR-interval values from all the RR-interval values in an epoch.

These feature use as an attributes into our multinominal logistic regression classifier.
We conducted experiments to obtain optimal cross validation for detecting apnea sleep disorder. As performance evaluation we use classification accuracy (CA), or confusion matrix.

A. K-fold

K-fold cross validation is one of the cross validation method which divide dataset into k datasets and the holdout method is iterated k times. The cross-validation K-Fold is the data resulted from the observation of the mentioned sub-groups to be K which denotes the number of sub-groups. A sub-group of data is used for validation and testing and the rest sub-group K – 1 is for training process. The mentioned process is repeated for K times with each sub-group used one time only. The yield of such calculation is the mean value to get the final value. The illustration of k-fold cross validation depicted on the figure 1. The advantage of this method is that all observed data is used either in training or testing process. The experiment conducted use 3, 5, 10, 15, and 20 folds cross validation.

B. Repeated Random Sub-sampling (Bootstrapping)

In the bootstrapping, the data was split into two groups, training and testing data with particular composition. The data compositions are constructed by drawing from the original data randomly. Then, n-percent of the data used as training data and the remaining used as testing data. The experiments used 50, 60, 70, 80 and 90 percent of data as training data.

VI. RESULT AND ANALYSIS

A. K-fold

The experimental result using various cross validation shows that higher fold gives a better accuracy result. The experimental result of k-fold cross validation depicted on the Figure 2. The highest classification accuracy was obtained for the 15 and 20 folds. The lowest classification accuracy obtained on the 3-fold cross validation. In k-fold cross validation, the data separated in k groups. The more groups formed the number of training and testing data are getting significantly different. This would results in the condition where the experiment use large number of data for training while for testing the system uses fewer data.

Confusion matrix could be used for further analysis. The confusion matrix of this experiment depicted on fig 4. Based on confusion matrix, the class B always obtains constant recognition in any k-fold experiments. It means that the recognition of class B could not enhanced in any fold election in cross validation. But for class A and N, it can be seen that the classification accuracy is increased when higher number of k is selected. But for the 15 and 20 folds we obtained the same classification accuracy. It can be analyzed using confusion matrix.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>THE CONFUSION MATRIX OF 3 (A), 5 (B), 10 (C), 15 (D), 20 (E) FOLD CROSS VALIDATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>3003</td>
</tr>
<tr>
<td>B</td>
<td>171</td>
</tr>
<tr>
<td>C</td>
<td>1094</td>
</tr>
</tbody>
</table>

Based on the confusion matrix, the correct prediction of class A in 15 fold is higher than 20 fold. Meanwhile, in the class N the correct prediction of 20 fold is higher than 15 fold. Hence, the composition of training and testing data in 15-fold is the best suited to parameter optimization of each attribute for logistic regression in class A. Meanwhile for the class B, the composition of training and testing data in 20 fold is the best.
B. Repeated Random Sub-sampling (Bootstrapping)

Bootstrapping experimental result shows as the percentage of data used for training is higher, the better accuracy result. The experimental results of bootstrapping depicted on the Figure 3. The highest classification accuracy is obtained when 80 percent of data are used for training meanwhile the lowest is obtained when 50 percent of data are used for training. The more data used in training phase, it would be increase the performance of system and vice versa. The more data training would give better data representation of the system results in higher classification accuracy.

However for the 90 percent of data used as data training has lower classification accuracy than 80 percent of data used as data training the more data training would bring an over-fitting in regression. Generally, increasing the proportion of data used for training after suitable model is found would only introduce an error or noise rather than a good data representation.

VII. CONCLUSION

Cross validation is one of the validation methods to predict the fit of a model to a prediction validation set when an obvious validation set is not available. One of the cross validation type is k-fold cross validation.

Based on the experimental result, the higher numbers of fold in k-fold cross validation would increase the classification accuracy even slightly. But after an optimal k is obtained, increasing the proportion of training data (by increasing k) will not give any difference. On the other hand, the bootstrapping experimental results show that up to a certain proportion, the higher n-percent data are used for training might increases the classification accuracy. But after this certain proportion is found, we can see that the classifier will be degraded due to over-fitting between the data and the model to be estimated. In addition, compared to k-fold cross validation, bootstrapping give a higher accuracy prediction because it gives a more accurate data presentation by introducing randomness in selecting the training and testing data.

VIII. FURTHER WORKS

The apnea disorder detection system is continuously becoming an interesting research topic that draws many attentions and more applications in the future. With the advanced of technology, many diseases could be detected and treated early. Other approaches can be studied further to increase the classification accuracy such as by using more precise features in ECG that contains the most information and by using other polysomnogram channels. In the future, we are planning to examine logistic regression not only to detect apnea but also sleep stages.


