Cross validation methods: Analysis based on diagnostics of thyroid cancer metastasis

Mariusz Rafało

Warsaw School of Economics, Institute of Information Systems and Digital Economy, ul. Madalinskiiego 6/8, 02-513 Warsaw, Poland

Received 8 February 2021; accepted 2 May 2021
Available online 8 May 2021

Abstract

Much of modern medical research are supported by statistics and machine learning techniques. For the reliability of these studies, it is important to measure the effectiveness of the developed models in making predictions on real medical data. Cross validation is a tool, which provides a principled method of measuring the effectiveness of models and comparing models with each other. This paper presents theoretical and empirical analysis of common cross validation methods. Using sample logistic regression model along with six different validation methods, comparative analysis has been conducted. The model was developed to determine the level of thyroglobulin hormone, thyroid cancer marker. The marker indicates an increased risk of thyroid cancer metastasis, and is commonly used in metastasis diagnosis. Whereas, literature does not provide a single threshold of thyroglobulin value, indicating an increased risk of metastasis. This paper confirms that one of the main reasons of this lack, is the high variance of the developed models. This research argues that the use of the validation technique also influences both: measures of the model quality and the threshold value. The results show a high discrepancy in the determination of threshold values. However, iterative methods (such as sampling and bootstrap) seem to produce more stable outcomes.

Keywords: Machine learning; Model validation; Monte Carlo validation; Cross validation; LOOCV; Thyroid cancer

1. Introduction

Contemporary medical diagnostics is largely based on the use of statistical tools and more often on machine learning techniques. Machine learning models in particular require independent testing before they can be used in practical diagnosis. A common method of evaluating a machine learning models is cross validation. The method is based on the assumption that the model is trained on a different dataset than is used for testing. The model identifies rules on one dataset (training), then these rules are validated on another dataset (test or validation dataset). Validation dataset contains information about real classification result, so model accuracy can be tested objectively.

The goal of this paper is to conduct a comparative analysis of most common methods of machine learning model validation. The analysis was performed on the example of an econometric logistic regression model. The purpose of the model classification was to identify a threshold value of thyroglobulin hormone in diagnosis of malignant thyroid cancer metastasis.

Thyroid cancer (especially papillary type) is the most common malignancy of the thyroid and has favorable long-term survival in most cases. However, the frequency of cervical lymph node involvement is 27% to 46%. The recurrence rate is 3% up to 30% during postoperative treatment [1]. Therefore, distinguishing lymph node metastasis from benign reactive lymph adenitis (which are frequent in neck area) is important, both to avoid unnecessary treatment and to expediently detect malignancy in patients with thyroid cancer [2]. Furthermore, metastasis in cervical lymph node from other (non thyroid) cancers is also relatively common [1]. Additionally, it is stated that, ultrasound examination alone is not enough to distinguish a metastatic nodule from benign one [3].

Selecting specific validation method implies accepting its limitations, such possibility of generalization or the level of variance. The key motivation to undertake the research is the need to reliably define rules for medical diagnosticians. The performed analysis concerns the diagnosis of lymph nodes
2.1. Model validation methods

The diagnosis of malignant thyroid cancer metastasis is typically conducted based on several information sources: the analysis of markers from the blood test during blood collection, washout fluid from fine needle biopsy (FNA) and the ultrasound image. The primary goal is to define threshold values for markers and ultrasound features that will allow diagnosticians to determine the level of cancer advancement and the risk of nodal metastasis. Ultrasound and fine needle aspiration biopsy are commonly used and trustworthy diagnostic tools for cervical metastasis of differentiated papillary thyroid cancer [4].

The key problem for diagnosis is to determine clear values of markers (e.g. thyroglobulin) that can be treated as malignancy indicator. Many methods, such as Akaike information criterion [5], resubstitution validation or the cross validation, can be used to verify accuracy of these values. Among these methods, cross validation is of most applications in chemometrics [6].

The methods of model validation can be sorted according to the number of model iterations performed. The more iterations in the method, the more divisions of the dataset into smaller ones and the more models built. Starting from single validation on the training set, through hold-out validation, $k$-fold cross validation, to leave-one-out validation and validations based on Monte Carlo sampling and bootstrapping. The article analyzes and compares the following validation methods:

- Resubstitution validation
- Hold-out cross validation
- $k$-fold cross validation
- Leave-one-out (LOOCV) cross validation (also referred as the Jackknife cross validation)
- Random sub-sampling cross validation
- Bootstrap cross validation

2. Literature review

This section presents key literature analysis. Relevant literature from multiple sources is referred for model validation and cross validation analysis. Validation of econometric or machine learning model is a procedure of evaluating the goodness of model performance against the real data. Usually, it is a different dataset from the one on which the model was trained. In the process of creating and implementing machine learning models, it is an important step because it gives the confidence that the model will properly classify real data.

2.1. Model validation methods

For a systematic review of the available methods, a number of parameters should be taken into account. Consider a dataset $N$, consisting of $n$ elements ($n$ equals the number of observations) and $k$ is the number of the subsets in validation process. Let $j$ be the number of iterations and $i$ define how many models are built during the validation process, predominantly $i = k = j$, but for bootstrapping methods more often $j > k$. Let $n_t$ be the size of the training set and $n_v$ is the size of the validation set. For cross validation, training cases are randomly selected. The number of cases assigned to the training dataset is most often defined as a percentage of the size of the entire dataset. The percentage ($r$) mostly equals 70% or 80%, therefore validation dataset size is 30% and 20% respectively. Using this notation, all validation methods can be divided into the following groups (see Table 1).

- Resubstitution validation. All observations are used to build the model, and all observations are used to validate the model. This approach leads to a situation where the model fits only the training data, and for real data it generates a significant bias. The error is defined as the resubstitution error (measured on training set). In this case, $n_t = n_v$, and moreover $n_t = n_v = n$.
- Hold-out validation. In order to avoid resubstitution error, it is common to divide the dataset into two mutually exclusive datasets: training dataset and test (validation) dataset. Only one model is build, so $k = i = 1$. As mentioned, the split proportion $r$ may range from 80%/20%, 70%/30% to 60%/40%. This method creates a risk that the split of the data does not correspond with the predicted class distribution (especially when the predicted phenomenon is relatively rare). To solve this risk, the dataset can be divided into equal instances of classes in both datasets (by using stratified sampling). For extremely rare predicted observations, the SMOTE (Synthetic Minority Over-sampling Technique) technique can be used. The technique is to generate artificial observations that are very similar to real ones [7].
- $k$-fold cross validation. In this method, $n_t = n - \frac{v}{k}$ and $n_v = n - n_t = \frac{v}{k$. $k$-fold cross validation divides the $n$ observations into $k$ mutually exclusive subsets of equal size. The process then leaves out one of the $k$ subsets as a test dataset and trains the model on the remaining subsets (treated as one dataset). This process is repeated $k$ times, leaving out one of the $k$ subsets each time. This method provides $k$ times validation, which can reduce limitations of hold-out validation. Number of models equals $k$ ($i = k$). The results of a $k$-fold cross validation are often summarized as a mean of the model scores. It is also good practice to include a measure of the variance of the model, such as the standard deviation or standard error [8]. The choice of $k$ is usually 5 or 10 [9], but there is no formal rule [10,11]. For smaller values of $k$, the difference in size between the training dataset and the validation subsets gets smaller. As this difference decreases, the bias of the technique becomes smaller [12]. The size of $k$ can range from 1 (which brings it to hold-out validation) to $n$, (it is then called leave-one-out cross validation).
- Leave-one-out cross validation (LOOCV). In this method, only one observation is used for testing, while other observations are used for training the model. In LOOCV $i = n$, while $n_v = 1$ and $n_t = n - 1$. All observations are used both for training and testing, which
This phenomenon is called variance–bias trade-off [15,16]. Low variance, and models with high variance have low bias. In general, models with high bias have higher variance, so the results will vary, depending on data used to train the model. In general, models with high bias have low variance, and models with high variance have low bias. This phenomenon is called variance–bias trade-off [15,16].

Choosing different values of $k$, $i$ and $n_i$ affect this trade-off. Larger values of $k$ or $n_i$ produce less biased models but also larger training sets are more similar between iterations, hence the risk of overfitting increases [17]. For $k$-fold cross validation, common $k$ values are 5 or 10, as these values have been shown empirically to suffer neither from excessively high bias nor from very high variance [18].

As indicated, each method has its own advantages and limitations. Resubstitution method fits a single classifier to the data, and applies this classifier to each data observation (in the same dataset). Resubstitution typically underestimates model error, while cross validation has the advantage of producing an effectively unbiased error estimate, but the estimate is highly variable [19]. LOOCV method removes each observation in each iteration, then constructs the classifier, and then computes whether this leave-one-out classifier correctly classifies the removed observation. When using cross validation, each observation gets tested exactly once, which seems methodically correct. However, cross validation only explores a few of the possible ways that data could have been partitioned. Random sub-sampling lets explore more possible partitions, though it is unlikely to get all of the combinations.

### 3. Methodology

The study was conducted as a series of simulations. Each simulation is an autonomous model validation by applying different method. The approach is consistent with Conway [20], who describes a series of statistical experiments as simulation. This implies a specific methodological, mathematical and notation rigor. Conway points to a number of issues related to the statistical approach, such as the possibility of comparing simulation results or the methods of measuring random error [20].

The purpose of each simulation is to build and validate logistic regression model. In the model, dependent variable is a classification of thyroid cancer metastasis (binary classification). Independent variable is thyroglobulin hormone level, measured in fine needle biopsy washout fluid (this variable is referred as Tg-FNA). Each simulation included the steps of building and validating the model. The model evaluation was performed using the receiver operating characteristic (ROC) curve. Additionally, the simulation determined Tg-FNA threshold, above which the risk of metastasis increases.

#### 3.1. Dataset

It is important that the dataset used for analysis is well-annotated. Especially, when there is a need to produce robust method for medical diagnosis. Getting to good quality dataset is a difficult task, due to the unavailability of reliable data and confidentiality of the patients’ demographic information [21].

Dataset that was used for analysis comes from The Nuclear Medicine and Endocrine Oncology Department of Maria Sklodowska-Curie Institute — Oncology Centre, located in Warsaw, Poland. The dataset includes variables related to the diagnostics of thyroid cancer metastasis. The dataset contains variables describing cancer markers and the ultrasound image features. Additionally, demographic variables, in particular gender and age of the patient, are also available. The collection includes data from 200 patients.
The predicted class concerns the presence of a malignant metastasis to the lymph nodes. It is a binary classifier that takes value 0 — for benign nodule and 1 for metastasis. Distribution of this variable is presented in Table 2.

Initially, two markers were analyzed. The analysis concerned thyroglobulin level from blood test (Tg) and thyroglobulin level from fine needle biopsy (Tg-FNA). It was verified whether the presence of a lymph node metastasis significantly differentiates the level of thyroglobulin. The analyses were carried out using the t-student test. For Tg value, \( p = 0.0302 \) (\( n = 198 \)), for Tg-FNA, \( p = 0.000 \) (\( n = 200 \)). For both markers \( p \)-value was lower than confidence level \( \alpha = 0.05 \). However, at the stage of model parameters estimation, it turned out that the Tg-FNA variable differentiates patients sample much better. Thus, only the Tg-FNA variable was included in the final model.

Due to the high differences in the levels of Tg-FNA (Table 3), a logistic regression model was applied. To simplify the analysis, only one variable was included in the model and the focus was on determining the most reliable threshold of Tg-FNA value.

Simple models, such as linear regression and logistic regression generally tend to have a higher bias and corresponding low variance. This observation is consistent with the purpose of the study to determine the effect of validation techniques on the classification results. The low variance of the model itself will help to obtain reliable results. Thus, the increased variance will result from the validation methods used and not from modeling techniques themselves.

The construction of the model was preceded by exploratory analysis. Selected descriptive statistics of Tg-FNA values are presented in Table 3, for benign and malignant metastatic cases. Distribution of patients by Tg-FNA levels is presented in Table 4.

### Table 2

<table>
<thead>
<tr>
<th>Lymph node condition</th>
<th>Number of patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benign lymph node (0)</td>
<td>140</td>
</tr>
<tr>
<td>Metastatic lymph node (1)</td>
<td>60</td>
</tr>
</tbody>
</table>

### Table 3

| Tg-FNA summary form metastatic and benign lymph nodes (\( p = 0.000, n = 200 \)). |
|---------------------------------|-------------------|
| Statistics                      | All patients      | Metastatic | Benign  |
| Mean                            | 493.54            | 1510.12    | 57.86   |
| Standard deviation              | 1082.27           | 1486.36    | 319.19  |
| Min                             | 0.04              | 0.45       | 0.04    |
| Max                             | 5000              | 5000       | 2500    |
| Median                          | 3.50              | 958.50     | 1.96    |

### Table 4

<table>
<thead>
<tr>
<th>Tg-FNA interval</th>
<th>Metastatic</th>
<th>Benign</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0.0-1]</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>(0.1-2]</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>(2-15]</td>
<td>3</td>
<td>59</td>
</tr>
<tr>
<td>(15-90]</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>(90-1000]</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>&gt;1000</td>
<td>30</td>
<td>3</td>
</tr>
</tbody>
</table>

The model parameters estimation process was carried out in accordance with the adopted validation method. The number of models built varied, as some methods assume validation in stages or iteratively. For each model, quality measures were determined based on the ROC curve analysis. The receiver operating characteristic curve as a tool for objective evaluation and comparison of classification models [9,15,22]. It is also a method commonly used in medical research [2,4,23]. The following model quality measures have been analyzed:

- TPR — true positive rate (sensitivity)
- TNR — true negative rate (specificity)
- PPV — positive predicted value (precision)
- NPV — negative predictive value

For validation methods that generate more models, the model quality measures were aggregated. For the ROC indicators (TPR, TNR, PPV and NPV) average values are presented. Also, in order to present variance of these measures, standard deviation and median are also included. The results of each simulation are presented in Tables 5 and 6.

### Table 5

<table>
<thead>
<tr>
<th>Method</th>
<th>TPR</th>
<th>TNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resubstitution</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>Hold out validation</td>
<td>0.87</td>
<td>0.89</td>
</tr>
<tr>
<td>5-fold cross validation</td>
<td>0.93</td>
<td>0.94</td>
</tr>
<tr>
<td>10-fold cross validation</td>
<td>0.99</td>
<td>0.91</td>
</tr>
<tr>
<td>LOOCV</td>
<td>0.88</td>
<td>0.94</td>
</tr>
<tr>
<td>Sub-sampling cross validation</td>
<td>0.94</td>
<td>0.93</td>
</tr>
<tr>
<td>Bootstrap cross validation</td>
<td>0.94</td>
<td>0.93</td>
</tr>
</tbody>
</table>

### Table 6

<table>
<thead>
<tr>
<th>Method</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resubstitution</td>
<td>0.84</td>
<td>0.97</td>
</tr>
<tr>
<td>Hold out validation</td>
<td>0.76</td>
<td>0.94</td>
</tr>
<tr>
<td>5-fold cross validation</td>
<td>0.86</td>
<td>0.97</td>
</tr>
<tr>
<td>10-fold cross validation</td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>LOOCV</td>
<td>0.88</td>
<td>0.96</td>
</tr>
<tr>
<td>Sub-sampling cross validation</td>
<td>0.85</td>
<td>0.98</td>
</tr>
<tr>
<td>Bootstrap cross validation</td>
<td>0.86</td>
<td>0.97</td>
</tr>
</tbody>
</table>

3.2. Cross validation

The model parameters estimation process was carried out in accordance with the adopted validation method. The number of models built varied, as some methods assume validation in stages or iteratively. For each model, quality measures were determined based on the ROC curve analysis. The receiver operating characteristic curve as a tool for objective evaluation and comparison of classification models [9,15,22]. It is also a method commonly used in medical research [2,4,23].

An important element of the analysis, resulting from the purpose of the study, was also the determination of the Tg-FNA threshold, supporting the diagnosis of thyroid cancer metastasis. For each validation method, a threshold value was determined. For the methods that generated more models, the average threshold values, along with standard deviations are presented (Table 7). The inclusion of the standard deviation indicates how stable the threshold values are generated by the given validation method.
4. Results and discussion

Previous studies of the Tg-FNA cut-off value report different threshold values, ranging from 1 ng/mL [2], 5 ng/mL [4] up to 100 ng/mL [1]. Due to the relatively large discrepancies in the results, the observed practice is to set the threshold to the mean of Tg-FNA, increased by 2 standard deviations [3]. In this study, the threshold value ranges from 10 to 71, depending on the validation method used. Relatively similar threshold values were obtained using methods with high numbers of iterations: sub-sampling (37, median = 11) and bootstrap (45, median = 12).

The models developed in this study report relatively high level of sensitivity and specificity (along with other ROC measures). TPR ranges from 0.87 (for hold out method) to 0.99 ± 0.04 (for 10-fold cross validation), TNR ranges from 0.89 (hold out method) to 0.94 (for LOOCV). All the various aspects of classifier training take place inside the validation loop. It has been proven, that all aspects of training a classifier, like feature selection, type selection or parameter tuning should take place inside each validation loop [24]. The conducted analysis meets the assumption: in each iteration, the relationships between the Tg-FNA and metastatic variables were measured. Model parameters were also estimated inside each loop.

The rule of thumb is to increase model complexity, thus decreasing bias and increasing variance, until bias has been minimized and before significant variance errors become evident [25]. This study supports this statement: the values of Tg-FNA threshold values are significantly different depending on the method used. At the same time, the model error measures (TPR, TNR and derivatives) show relatively little changes.

5. Conclusions

The article presents a comparative analysis of the most popular methods of model validation. The analysis was performed with the focus on stability of the results. The conclusion of the analysis is that the choice of a validation technique has a significant impact on the cut-off (threshold) value (Tg-FNA threshold in this study). This is particularly important in medical practice, where the identified threshold values are used to support diagnoses.

The work has significant practical implications for medical diagnostics and for the analysis of cancer markers. Detailed analysis was performed to identify metastases after a history of thyroid cancer, however the conclusions can be successfully generalized to other types of cancer. The practical application of the simulation results is the confirmation of the necessity of using several methods of model validation, especially by controlling the threshold values along with ROC curve measures.

In weighing the results of this study, several limitations should be considered. First, the dataset was relatively small (n = 200). However, medical research and practice show that the research of this field often rely on relatively small number of observations (e.g. 43 cases [1]). From an analytical point of view, a larger dataset size would allow the simulation to be carried out in a wider context (for example, by including a larger k-value in the k-fold cross validation). Secondly, to illustrate the phenomenon, only one analytical technique was used (logistic regression model) and the model based on only one explanatory variable (Tg-FNA hormone level).

The directions of further research result from the limitations of this study. Further research should then focus on developing specialized methods of selecting cut-off points for cancer markers. The lack of consistency in the validation of the models may be one of the reasons for large discrepancies in thresholds reported in literature. Particular attention should be paid to research on the selection of the validation method for the applied data analysis technique.

Declaration of competing interest

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.

Acknowledgment

The author would like to express appreciation and thanks to Professor Marek Dedecjus at Maria Sklodowska-Curie National Research Institute of Oncology, Poland, for support and help in sharing the dataset.

References