Informatics and Technology in Clinical Care and Public Health
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210876

A-Test Method for Quantifying Structural Risk and Learning Capacity of Supervised Machine Learning Methods

Arash GHAREHBAGHI[a,1] and Ankica BABIC[a,b]
[a] *Department of Biomedical Engineering, Linköping University, Sweden*
[b] *Department of Information Science and Media Studies, University of Bergen, Norway*

Abstract. This paper presents an original method for studying the performance of the supervised Machine Learning (ML) methods, the A-Test method. The method offers the possibility of investigating the structural risk as well as the learning capacity of ML methods in a quantitating manner. A-Test provides a powerful validation method for the learning methods with small or medium size of the learning data, where overfitting is regarded as a common problem of learning. Such a condition can occur in many applications of bioinformatics and biomedical engineering in which access to a large dataset is a challengeable task. Performance of the A-Test method is explored by validation of two ML methods, using real datasets of heart sound signals. The datasets comprise of children cases with a normal heart condition as well as 4 pathological cases: aortic stenosis, ventricular septal defect, mitral regurgitation, and pulmonary stenosis. It is observed that the A-Test method provides further comprehensive and more realistic information about the performance of the classification methods as compared to the existing alternatives, the K-fold validation and repeated random sub-sampling.

Keywords. A-Test method, structural risk, learning capacity, heart sounds

1. Introduction

Artificial Intelligence (AI), as an advancing context, is creating a significant influence on different aspects of social sustainability including healthcare. AI-based tools are consistently becoming part of the healthcare system towards providing better healthcare for all the individuals of a society, where supervised machine learning methods serve as a central part for making the appropriate medical decision [1–5]. The performance of such machine learning methods is critically important, sometimes with vital value, as a mistaken error can lead to incorrect patient management. It is, therefore, crucial for any machine learning method to be properly trained, especially when it comes to medical applications [6]. In general, there are two main circumstances in the training of the machine learning methods, small-medium size and large size of the learning data. The main challenge for the large size data is to avoid biased training on any group. On the other hand, the small-medium size data is more likely to face overfitting, which affects the reproducibility of the results [6–7]. Small-medium size data is seen in many

[1] Corresponding Author, Arash Gharehbaghi. Department of Biomedical Engineering, Linköping University, Sweden: arash.gharehbaghi@liu.se

applications of biomedical engineering and bioinformatics, where the data acquisition is problematic, particularly, under the limitations recommended by the new adaptation of Good Clinical Practices and the codes of the General Data Protection Regulation (GDPR). Extensive attention has been paid by the AI community to improve learning methods, whereas validation has been by far less investigated [8–9]. Repeated random sub-sampling and K-Fold, are considered as the two existing validation methods to evaluate the accuracy of supervised classification methods. These methods have been commonly employed by the researchers for validation purposes. Both of these two methods lack the capability of proving realistic and pervasive evaluation. In this paper, we introduce novel capabilities of our original validation method, the A-Test method, in estimating the learning capacity of any supervised classification method [10]. A study is performed for two different classification methods trained for a demanding clinical application, and the results are illustrated and compared.

2. Materials

Heart sound signals were recorded from the referrals to the Children Medical Centre of Tehran, using an electronic stethoscope of WelchAllyn Meditron Analyzer in conjunction with a portable computer. All the referrals underwent echocardiography, and the study was approved by the appointed ethics committee and was conducted according to the Good Clinical Practice. All the referrals or their legal guardians gave their informed consent to participate in the study. The patient population is listed in Table 1.

Table 1. Patient population of the study.

Heart Condition	Number of Patients	Age Range (years)
Aortic Stenosis	15	1–8
Mitral Regurgitation	15	4–8
Normal without murmur	30	4–15
Pulmonary Stenosis	15	1–10
Ventricular Septal Defect	25	1–9

3. Methods

3.1. Structural Risk and Learning Capacity

Structural risk of a classification method is defined as instability of performance measure of the classification method when the method is tested by a dataset out of the training data. The learning capacity of a classification method is defined as the capability of improvement in the performance of a classification method when the method is trained by a broader set of training data.

3.2. The A-Test Method

The A-Test method is based on using k-fold validation method for different values of *k*. In k-fold validation, the validation dataset is divided into *k* partitions with almost equal length. One partition is used for testing and the rest for training the classification method. This procedure is repeated k times with one partition is used only once for testing. The

A-Test employs the k-fold validation with different values of k $(k=2,\ldots,K_{max})$, and the classification error is calculated for each k-value:

$$\Gamma_{\mathrm{M}} = 100\ \frac{\sum_{k=2}^{K_{max}} \Gamma_{M,k}}{K_{max}-1} \tag{1}$$

where $\Gamma_{M,k}$ is the classification error of the classification method M, and k is the fold value for validation, called validation index. K_{max} is less than the minimum group size of the validation data. For a classification method, the percentage of the relative span of the classification rate, $\psi_{M,k}$, is employed as an indication of the learning capacity.

$$\mathrm{P}_{\mathrm{M}} = 100\ \frac{\max\limits_{k} \psi_{M,k} - \min\limits_{k} \psi_{M,k}}{\max\limits_{k} \psi_{M,k}} \tag{2}$$

3.3. The classification methods

The A-Test method is employed to compare the structural risk of two different learning methods for classifying PCG signals, a deep time growing neural network (DTGNN) and a hidden Markov model (HMM), whose technical details are found in [10] and [11–12], respectively. Both the DTGNN and HMM are trained to learn the pathological characteristics of the signal, caused by aortic stenosis.

4. Results

Figure 1 shows a variation of the classification error due to the k-value. The descriptive statistics of the classification error, as well as the learning capacities are listed in Table 2.

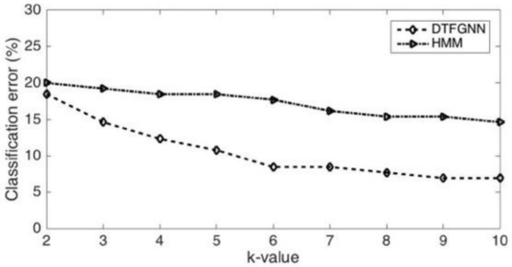

Figure 1. Variation of the classification error with respect to k-value for the two learning methods.

It is observed that the DTGNN provides a lower classification error, especially for higher k-values with a lower structural risk. Interestingly, for lower k-values, the two methods do not substantially differ in their performance.

Table 2. Descriptive statistics of the classification error for the two learning methods, the deep time growing neural network (DTGNN) and the hidden Markov mode (HMM).

Statistics	DTGNN	HMM
Average (%)	10.51	17.27
Minimum (%)	6.92	14.62
Maximum (%)	18.46	20.00
Median (%)	8.46	17.69
Learning Capacity (%)	14.2	6.7

5. Discussion

There are mainly two alternatives to the A-Test for exploring the structural risk of a classifier: the k-fold and the repeated random sub-sampling methods, both cannot provide an understanding of the learning capacity. Furthermore, it might lead to incorrect comparison for certain values of k, as was the case for the DTGNN and HMM with $k=2$. The learning capacity for the DTGNN and HMM is 14.2% and 6.7%, respectively, showing a higher capacity for the DTGNN because of the deep architecture.

6. Conclusions

This paper suggested the A-Test, as a powerful validation method. A-Test method provides means for validating not only the performance of a classifier but also the structural risk and the learning capacity of classification methods, by validating the classification methods' different ratios of training/test data and quantifying the results. This aspect cannot be seen in other alternatives, K-Fold validation and repeated random sub-sampling, which can potentially lead to an inappropriate validation result.

References

[1] Gharehbaghi A, Lindén M, Babic A. An Artificial Intelligent-Based Model for Detecting Systolic Pathological Patterns of Phonocardiogram based on Time-Growing Neural Network. Applied Soft Computing. 2019;83:105615.

[2] Gharehbaghi A, Ask P, Lindén M, Babic A. A novel model for screening aortic stenosis using phonocardiogram. In16th Nordic-Baltic Conference on Biomedical Engineering; 2015; Springer, Cham; p. 48-51.

[3] Gharehbaghi A, Sepehri AA, Linden M, Babic A. Intelligent Phonocardiography for Screening Ventricular Septal Defect Using Time Growing Neural Network. InICIMTH; 2017 Jul; p. 108-111.

[4] Gharehbaghi A, Lindén M, Babic A. A decision support system for cardiac disease diagnosis based on machine learning methods. Stud Health Technol Inform. 2017 Jan 1;235:43-7.

[5] Gharehbaghi A, Lindén M. An internet-based tool for pediatric cardiac disease diagnosis using intelligent phonocardiography. InInternational Internet of Things Summit. 2015 Oct 27; Springer. Cham; p. 443-447.

[6] Jain AK, et al. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2000;22(1):4-37.

[7] Gharehbaghi A, Babic A. Structural Risk Evaluation of a Deep Neural Network and a Markov Model in Extracting Medical Information from Phonocardiography. Studies in Health Technology and Informatics. 251:157-160.

[8] Gharehbaghi A, et al. An Intelligent Method for Discrimination between Aortic and Pulmonary Stenosis using Phonocardiogram, in World Congress on Medical Physics and Biomedical Engineering; 2015; Springer International Publishing: Cham. p. 1010-1013.

[9] Gharehbaghi A, et al. A Hybrid Machine Learning Method for Detecting Cardiac Ejection Murmurs. in EMBEC & NBC. Singapore; 2018; Springer Singapore; p. 787 – 790.

[10] Gharehbaghi A, Lindén M. A deep machine learning method for classifying cyclic time series of biological signals using time-growing neural network. IEEE transactions on neural networks and learning systems. 2017 Oct 12;29(9):4102-15.

[11] Gharehbaghi A, Ask P, Babic A. A pattern recognition framework for detecting dynamic changes on cyclic time series. Pattern Recognition. 2015;48(3):696-708.

[12] Gharehbaghi A, et al. A Hybrid Model for Diagnosing Sever Aortic Stenosis in Asymptomatic Patients using Phonocardiogram, in World Congress on Medical Physics and Biomedical Engineering; 2015; Springer International Publishing: Cham; p. 1006-1009.