Journal of Modern Technology and Engineering
Vol.6, No.3, 2021, pp.242-246

COMPARISON OF MACHINE LEARNING TECHNIQUES AND CLASSICAL STATISTICAL MODELS

Galip Savas Ilgi[1], **Eser Gemikonakli**[2*] , **Yoney Kirsal Ever**[3]

[1]Information Systems Engineering, Faculty of Engineering, Near East University, Nicosia, Mersin 10, Turkey
[2]Department of Computer Engineering, Faculty of Engineering, University of Kyrenia, Girne, Mersin 10, Turkey
[3]Software Engineering, Faculty of Engineering, Near East University, Nicosia, Mersin 10, Turkey

Abstract. In this study, the idea of using statistical hypothesis tests to analyse the performance of machine learning classifiers on a single data set was discussed. The nature of the data, rather than the learning technique, is of primary importance. Here, the naive application of statistical hypothesis tests can lead to misleading results. It may be better to use McNamer's statistical hypothesis test to analyse the performance of the machine learning classifiers and choose a final model. The data regarding the sex and age information of Covid-19 patients were collected. The analysis will be carried out to find the best model to predict the transmission risk between the groups as further study.

Keywords: SPSS, Machine learning, McNamers test, Mann–Whitney U test.
Corresponding author: Eser Gemikonakli, Department of Computer Engineering, Faculty of Engineering, Girne, Mersin 10, Turkey, e-mail: *eser.gemikonakli@kyrenia.edu.tr*
Received: 12 September 2021; Revised: 24 October 2021; Accepted: 2 November 2021; Published: 30 December 2021.

1 Introduction

Estimating the consequences of Covid-19 transmission risk from the available data has become important in health research and health management. It is usually evaluated by taking these data into account for gender and age classification. Such data rely on statistical models such as descriptive statistics and the Mann-Whitney U test model. Most such applications are based on the belief that there are relatively few important variables and that careful selection of these variables is key to the successful performance of models for outcome prediction.

Owning to the fact that more than three decades, computational methods have been increasing in use of mutli disciplinary fields, especially in health care systems. Estimating health outcomes from available data is an important challenge in health research and health management. It is usually evaluated by calculating scores/indices for risk classification (Singh et al., 2002). Traditionally, such scores are based on statistical models such as logistic regression (known as descriptive statistics) and the Mann-Whitney U test model (McKnight, 2010), (Singh et al., 2002). Most such applications are based on the belief that there are relatively few (though often unclear) important variables (risk factors) and that careful selection of these variables is key to the successful performance of models for outcome prediction.

However, such variables usually typically interact with each other in a complex, and unknown way and therefore they are often excluded from predictive models (Singh et al., 2002). As stated above, more than two decades, machine learning algorithms and techniques based on machine

learning (such as neural networks, support vector machines) have become available and they have been applied in different multidisciplinary fields (Ever et al., 2019). These approaches are based on inductive inference rather than classical statistics (Vapnik, 2000). Machine learning algorithms are not commonly found in statistical software packages, and even if they are, their applications often require skills that are outside of the usual experience of biostatisticians. Recently, in some researches different learning techniques with "classical" statistical algorithms are discussed and compared (Knuiman et al., 1997). However, such comparisons are often scarce, and only a small number of techniques have been investigated in a limited number of datasets (Vann et al., 2002).

In "classical" statistics, one of the most popular test method is the Mann–Whitney U test (which also known as the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that, for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X (Mann et al., 1947). In other words, the Mann-Whitney U test is used to compare whether there is a difference in the dependent variable for two independent groups. It compares whether the distribution of the dependent variable is the same for the two groups and therefore from the same population (Mann et al., 1947). Nevertheless, in many software packages, including most popular and effective ones such as SPSS, R, and Python, the Mann–Whitney U test (of the hypothesis of equal distributions against appropriate alternatives) has been poorly documented (Bergmann et al., 2000). Some packages incorrectly treat ties or fail to document asymptotic techniques (e.g., correction for continuity) (Bergmann et al., 2000). Additionally, it is found out that there are no existing machine learning algorithms applied to this hypothesis test model.

Furthermore, McNemar's test is a statistical test used on paired nominal data (Sun et al., 2010). It is applied to $2\Gamma - 2$ contingency tables with a dichotomous trait, with matched pairs of subjects, to determine whether the row and column marginal frequencies are equal (that is, whether there is "marginal homogeneity"). The commonly used parameters to assess a diagnostic test in medical sciences are sensitivity and specificity (Sun et al., 2010). Sensitivity is the ability of a test to correctly identify the people with disease. Specificity is the ability of the test to correctly identify those without the disease. Now presume two tests are performed on the same group of patients. And also presume that these tests have identical sensitivity and specificity (Sun et al., 2010).

Owning to the fact that considering the improvements in the artificial intelli- gence in last two decades are affecting the multi-disciplinary fields electronically. Within these developments many different and interesting proposals, frameworks and models are available in the literature. Especially, back propagation neural networks (BPNN) with prediction are highly acceptable and deployments have been increased especially in the fields of healthcare systems, finance and banking, agriculture, petroleum and gas and many more (Ever et al., 2019). Another example is provided in the same study stating that, many different discipliners have started to use BPNN prediction algorithms where the authors provided a new approach from the literature where it suggested and improved for training efficiency of BPNN algorithms (Ever et al., 2019). In this proposed algorithm modifies the gradient based search direction by introduing the value of gain parameter in the activation function. The proposed algorithm is generic and could be implemented in almost all gradient based optimization processes, because algorithm results enhanced the computational efficiency of training process. Their suggested algorithm's robustness and effectiveness are tested on a heart disease data set and they are shown by comparing convergence rates and gradient descent methods respectively (Ever et al., 2019).

Apart from BPNN models and algorithms, data mining techniques, especially Iterative Dichotomiser 3 (ID3) decision tree algorithms are widely used in liter- ature for prediction (Ever et al., 2019). Decision tree uses divide and conquer technique for the basic learning strategy (Ever et al., 2019). According to literature review searches in Demsar (2006), data mining is

the extraction of concealed prescient data from substantial databases, where these techniques understood future patterns and practices, in order to make proactive, learning driven choices. It is clearly stated that data mining processes work efficiently when a large amount of data is available (known as big data) (Ever et al., 2019). In data mining algorithms, grouping guideline is a procedure that arranges estimations of target variables from estimations of attributes of variables. The main objective of data mining is prediction. Existing and most applied machine learning techniques, can be listed as Back-propagation (BP) Learning Algorithm, Radial Basis Function Neural Network (RBFNN), Support Vector Regression (SVR) and Decision Tree Regressor (DTR) (Ever et al., 2019). Backpropagation is one of the most popular learning algorithms in neural net- works both for classification and prediction problems. It updates weights of each neuron using gradient descent algorithm and different kinds of stopping criteria can be used such as error level, iteration number etc. for the convergence of neural network. Radial Basis Function Neural Networks uses radial basis functions as an activation function of hidden neurons. It acts as universal approximation on sets and its' strength areas are function approximation, time series prediction and classification. Support Vector Regression is a form of Support Vector Machines to accept real value outputs instead of binary numbers. There are different types of SVR as linear and non-linear, and they can be used with different kernel functions such as polynomial, Gaussian radial basis function etc. Primarily, Decision Trees are proposed for the classification of data by using a divide-and-conquer strategy until final leaf. Then, they are modified to be used in regression models and, their simplicity and efficiency with large number of variables and cases make them popular for prediction problems.

The aim of our study is to compare the performance of machine learning technique (including those based on "classical" statistical models) with the help of the available data: that is, to compare statistical data with SPSS and statistical information for machine learning.

2 Models and Applications

Comparing machine learning methods and selecting a final model is a common operation inapplied machine learning (Ever et al., 2019; Demsar, 2006).

Models are commonly evaluated using resampling methods like k-fold cross-validation from which mean skill scores are calculated and compared directly. Although simple, this approach can be misleading as it is hard to know whether the difference between mean skill scores is real or the result of a statistical fluke. Statistical significance tests are designed to address this problem and quantify the likelihood of the samples of skill scores being observed given the assumption that they were drawn from the same distribution. In Demsar (2006) the author stated that, although there is not a significant proof statistical hypothesis testing can improve both your confidence in the interpretation and the presentation of results during model selection (Demsar, 2006).

Mann-Whitney U and McNemar Tests were applied to each data set. Samples were selected according to age groups and gender. And all calculations were repeated in accordance with these tests, allowing us to check the reproducibility of the contagion risk assessment in different subsamples and calculate the means and standard deviations for each month. Comparisons of the mean values obtained from each model were made using Analyzes of Variability (SPSS). Models were performed using both SPSS and PYTON software for simulations and statistical analysis.

3 Materials and Methods

In statistics, assumption of a statistical test is called null hypothesis and in order to decide whether or not to accept or reject the null hypothesis, statical measures are calculated and interpreted. Generally, a statistical hypothesis test for comparing samples quantifies how likely

it is to observe two data samples on a same distribution given the assumption (Demsar, 2006). In the case of selecting models based on their estimated skill it is important to know whether there is a real or statistically significant difference between the two models.

Comparing machine learning models via statistical significance tests imposes some expectations, Therefore the types of statistical tests that can be used are listed as follows (Demsar, 2006);

- Skill Estimate: In this test, a specific measure of model skill must be chosen. This could be classification accuracy (a proportion) or mean absolute error (summary statistic).

- Repeated Estimates: In repeated estimates, a sample of skill scores is required in order to calculate statistics. The repeated training and testing of a given model on the same or different data will impact the type of test that can be used.

- Distribution of Estimates: The sample of skill score estimates will have a distribution, such as Gaussian. This will determine whether parametric or nonparametric tests can be used.

- Central Tendency: In this test, model skill will often be described and compared using a summary statistic variable such as a mean or median.

The results of a statistical test are often a test statistic and a p-value. Both of these results can be used in the presentation of the results in order to quantify the level of confidence or significance in the difference between models (Demsar, 2006). This allows stronger claims to be made as part of model selection than not using statistical hypothesis tests.

4 Discussions

It should be noted that there is no silver bullet when it comes to choosing a statistical significance test for model selection in applied machine learning (Demsar, 2006).

Author of the research article Demsar (2006) stated that some of the researches showed that in order to select the appropriate statistical significance tests for model selection in machine learning several approaches are observed and discussed. The most preferred ones are listed as McNemar's test or $5\Gamma-2$ Cross-Validation, use of a Nonparametric Paired Test, use of estimated statistics.

Twenty-year long recommendations of McNemar's test for single-run classification accuracy results and $5\Gamma-2$-fold cross-validation with a generally modified paired Student's t-test. Also, additional correction to the Nadeau and Bengio test statistic can be used with $5\Gamma-2$-fold cross validation or $10\Gamma-10$-fold cross-validation as recommended by Weka developers. One difficulty with using the modified t statistic is that there is no out-of-the-box application (for example in SciPy) that requires the use of third-party code and the risks it entails. You may have to implement it yourself.

The availability and complexity of a chosen statistical method is an important consideration, as is well noted in Gitte Vanwinckelen and Hendrik Blockeel's 2012 paper, "Estimating Model Accuracy with Ten Predicted Cross-validation"(Vanminckelen et al., 2012). These methods have been carefully designed and compared to previous methods. Although it has been shown to improve in various ways, they suffer from the same risk as previous methods, meaning that the more complex a method is, the higher the risk that researchers will misuse it or misinterpret the result.

Instead of statistical hypothesis testing, estimation statistics such as confidence intervals can be calculated. These would suffer from similar problems where the independence assumption is violated given the resampling methods by which models are evaluated. Statistical methods such

as bootstrapping can be used to calculate defensible nonparametric confidence intervals, which can be used both to present results and to compare classifiers. This is a simple and effective approach that you can always rely on and is generally recommended.

References

Bergmann, R., Ludbrook, J. & Spooren Will, P.J.M. (2000). Different Outcomes of the Wilcoxon–Mann–Whitney Test from Different Statistics Packages. *The American Statistician*, *54*(1), 72-77.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, *7*, 1-30.

Ever, Y.K., Dimililer, K. & Sekeroglu, B. (2019). Comparison of machine learning techniques for prediction problems. In *Workshops of the International Conference on Advanced Information Networking and Applications*, 713-723.

Knuiman, M.W., Vu, H.T. & Segal, M.R (1997). An empirical comparison of multivariable methods for estimating risk of death from coronary heart disease. *J. Cardiovasc. Risk*, *4*(2), 127-34.

Mann, H.B. & Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, *18*(1), 50-60.

McKnight, P.E. & Najab, J. (2010). Mann-Whitney U Test. *The Corsini encyclopedia of psychology*. 1-1

Vanwinckelen, G. & Blockeel, H. (2012). On estimating model accuracy with repeated cross-validation, Gitte Vanwinckelen and Hendrik Blockeel. Proceedings of the 21st Belgian-Dutch Conference on Machine Learning, 39-44.

Singh, M., Reeder, G.S., Jacobsen, S.J., Weston, S., Killian, J.& Roger, V.L. (2002). Scores for postmyocardial infarction stratification in the community. *Circulation*, *106*(18), 2309-14.

Sun, X. & Yang, Z. (2010). Generalized McNemar's Test for Homogeneity of the Marginal Distributions. SAS Global Forum 2008 Statistics and Data Analysis Paper 382-2008. 39-44.

Vann, G.T., Suykens, J.A.K, Baesens, B., Viaene, S.,Vanthienen, J., Dedene, G.,De Moor, B.& Vandewalle, J. 2002. Benchmarking Least Squares Support Vector Machine Classifiers. *Neural Computation*, *15*, 1115-47.

Vapnik, V.N. (2000). *The Nature of Statistical Learning Theory*. New York: 2nd ed. Springer.