Multivariate two-sample hypothesis testing through AUC maximization for biomedical applications

Ioannis Bargiotas[*◇] Argyris Kalogeratos[*] Myrto Limnios[*]
Pierre-Paul Vidal[+] Damien Ricard[#] Nicolas Vayatis[*]

[*]Centre Borelli, Université de Paris - CNRS - SSA - ENS Paris-Saclay, France
[+]School of Automation, Hangzhou Dianzi University, Zhejiang, 310018, China
[#]Neurology Department HIA Percy, Service de Santé des Armées, Clamart, France

ABSTRACT

Clinical datasets usually carry numerous features (biomarkers, characteristics, etc.) concerning the examined populations. This fact, although beneficial, challenges the statistical analysis via standard univariate approaches. In the two-sample setting, the majority of the clinical studies evaluate their assumptions relying on a variety of available univariate tests, such as the Student's t-test or Mann-Whitney Wilcoxon. We developed an *easy-to-use-and-interpret* non-parametric two-sample hypothesis testing framework (ts-AUC) particularly using machine learning and the AUC maximization criterion. We test and verify the effectiveness of ts-AUC in real data containing posturographic features of Parkinsonian patients (PS) with and without history of falling.

KEYWORDS

Machine learning, AUC maximization, multivariate two-sample hypothesis tests, multiple testing, Parkinson's disease.

ACM Reference Format:
Ioannis Bargiotas[*◇] Argyris Kalogeratos[*] Myrto Limnios[*] and Pierre-Paul Vidal[+] Damien Ricard[#] Nicolas Vayatis[*]. 2020. Multivariate two-sample hypothesis testing through AUC maximization for biomedical applications. In *11th Hellenic Conference on Artificial Intelligence (SETN 2020), September 2–4, 2020, Athens, Greece.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3411408.3411422

1 INTRODUCTION

Clinical research often needs to find the significant differences between two groups of individuals (e.g. patients vs. healthy subjects). Researchers usually compute several features using signal processing and data mining techniques, and evaluate their usefulness relying on a variety of available univariate tests, such as the Student's t-test. Typically, multiple univariate tests are applied consecutively in order to find the statistically significant features. The aforementioned multiple testing scheme has been part of a well-known scientific debate [8], mainly criticized for the increased probability of reporting a false-positive finding [8]. Thus, many biostatisticians recommend to disclose all the analyses that have been done, not only the significant ones. The violation of this recommendation and the regular misuse of those tests [15] combined with the relatively small available cohorts, may lead to false conclusions and as a consequence to a significant lack of clinical consensus or at least delay in reaching it. Well-known adjustments have been proposed in order to limit the aforementioned probability of a false-positive finding, such as Bonferroni correction, although they have been reported as conservative compromises due to the significant increase of the probability for false-negative output [8].

The machine learning community has recently made significant progress in this topic [4, 10], especially related to the design of appropriate criteria for the characterization of the ranking performance and/or meaningful extensions of the empirical risk minimization approach to this framework [1, 5]. In many of these efforts, the well-known criterion of the area under the ROC curve (AUC) is considered as the gold standard for measuring the capacity of a scoring function to discriminate groups of populations [16]. Unfortunately, to the best of our knowledge, these novel advancements remain largely unexploited by the clinical communities.

The study's objective is to propose an *easy-to-use-and-interpret* two-sample hypothesis testing framework, orienting to clinical research. We propose a new variation of a multivariate two-sample test through AUC maximization [16]. Its effectiveness is tested using a multidimensional dataset that comes from the postural control assessment of patients with Parkinsonian syndromes (PS).

A standard way to assess postural control is using a force platform. A force platform records the displacement of the center of pressure (CoP) applied by the whole body during a session of measurement, while the individual stands upon it and follows the clinician's instructions. It has been shown that CoP trajectories (also called statokinesigrams) reflect individuals' postural impairment when special acquisition protocols are followed [13, 14]. The dataset we use includes two groups: fallers (PS_F) and non-fallers (PS_{NF}). This work highlights the benefits that one can have by using such kind of two-sample analysis in the presence of multiple features, and demonstrate the contradicting conclusions that a traditional statistical analysis might have had.

2 THE MULTIVARIATE TS-AUC TEST

Prior work. A key element of [16] is that the empirical AUC may be viewed as the Mann-Whitney statistic of a multivariate sample. They investigate the theoretical basis of how the rank-based test approach for homogeneity testing between two samples

[◇]Corresponding author. e-mail: ioannis.bargiotas@cmla.ens-cachan.fr

SETN 2020, September 2–4, 2020, Athens, Greece

ACM ISBN 978-1-4503-8878-8/20/09…$15.00
https://doi.org/10.1145/3411408.3411422

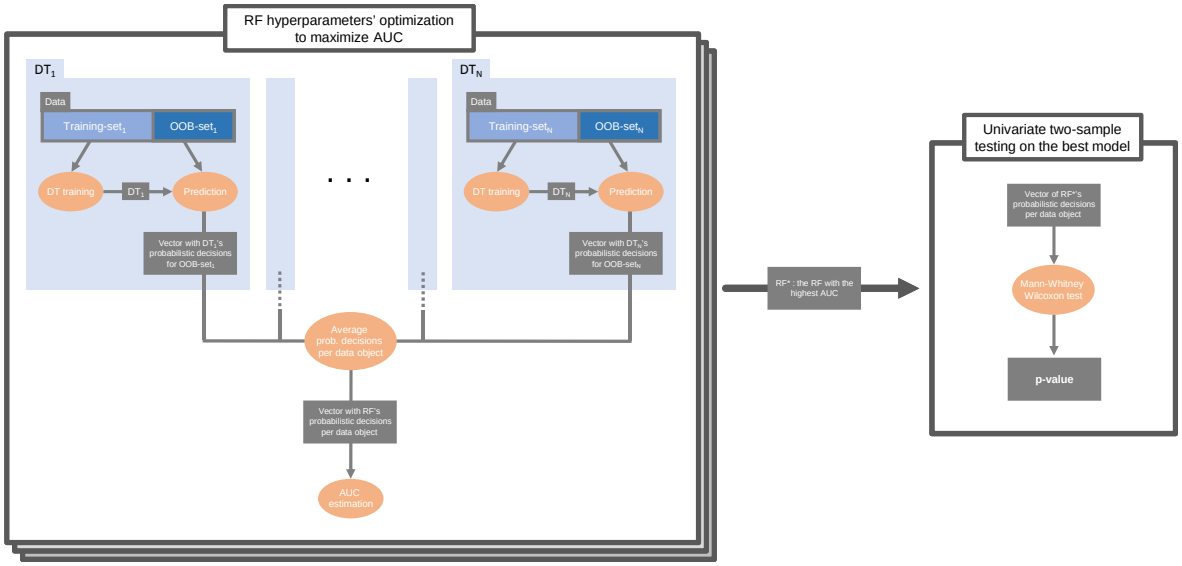

Figure 1: Scheme of the ts-AUC algorithm. In order to find the AUC* (maximal AUC), numerous Random Forests (RFs) are developed. For RF* with AUC*, the univariate Mann-Whitney Wilcoxon test is applied on the whole population's average positive posterior probabilities.

can be extended to a multidimensional setting. They propose a two-stage testing method based on data splitting. A nearly optimal scoring function in the AUC sense is first learned (greedy approach) from one half-sample. The remaining half-sample is then ranked according to the first stage's scoring function and a univariate Mann-Whitney Wilcoxon (MWW) two-sample test is applied.

Our proposed variation. The ts-AUC test, is based on a bootstrap aggregation, in particular over a random forest (RF) [3]. Therefore, in the development of each decision tree (DT), only a part of the whole dataset does participate (in-bag) while the other part is left out (out-of-bag, or OOB). The OOB subset is used as test-set for the particular DT. In our approach, instead of the originally proposed testing method based on data splitting, we used the predictions of the OOB population [7]. Every time an individual is part of an OOB set, the corresponding DT outputs the probability of being a PS_F or a PS_{NF}. This is computed as the fraction of individuals of the positive class (fallers) in the tree leaf where each individual reaches. Thus, his/her final score is given by the average of the posterior probabilities over the trees he/she was part of the OOB set (see Fig. 1).

The averaged posterior probabilities (P) of the positive class (fallers) are used in order to compute the Mann-Whitney U-test statistic, denoted by U as proposed in the theoretical work of [16] . The empirical AUC for the chosen hyper-parameters is given by $\frac{U}{N_{\text{fallers}} \cdot N_{\text{non-fallers}}}$. Briefly, the null hypothesis, H_0, and the alternative one, H_1, are expressed as:

$$\text{“}H_0 : \text{AUC}^* = \frac{1}{2}\text{”} \quad vs. \quad \text{“}H_1 : \text{AUC}^* > \frac{1}{2}\text{”}. \tag{1}$$

When searching for the empirical AUC* (i.e. the maximal AUC), the hyper-parameters with respect to which we need to optimize are the leaf-size LS and the number of features per tree M (see a comment about the computational cost in the Appendix file - www.bargiotas.com/material). We avoid a greedy approach and use instead a Bayesian optimization process. The averaged posterior probabilities of the Star Model, where AUC = AUC*, are used to compute the scoring function (and the p-value) through a univariate MWW test on the whole available dataset. Fig. 1 and Alg. 1 provide a schematic and algorithmic view of our statistical testing process that uses a bootstrap aggregation over an RF.

Feature importance. Additionally, the proposed algorithmic modifications allow the assessment of the importance of each feature to the ts-AUC final decision. We followed the procedure proposed in [9] for interpretation purposes, in order to identify all the important features, even some of those which are redundant/collinear. Briefly, we computed the AUC of the OOB (AUC_{OOB}) of RFs starting from the most important feature (see OOB feature importance by feature permutation [3]), adding progressively all the others in descending importance order. The best model is the smallest model (less features) with an AUC_{OOB} higher than the maximum AUC_{OOB} reduced by its empirical standard deviation (based on 20 runs).

3 EXPERIMENTS

Dataset. Our dataset comes from the Neurology department of the HIA, Percy hospital (Clamart, France), and includes 123 PS patients (36 women, 24/99 fallers/non-fallers, 78.7 ± 5.4 years-old – see Tab. 1 in the Appendix (www.bargiotas.com/material) for more details). Following the acquisition protocol, patients removed their shoes and maintained an upright position on a force platform keeping their arms at the sides. The CoP trajectory was recorded twice, with eyes open and eyes closed, for 25 seconds[1].

Statokinesigrams were acquired using a Wii Balance Board (WBB) (Nintendo, Kyoto, Japan), a suitable and convenient tool for the clinical setting [12]. Since the WBB records the CoP trajectories at non-stable time resolution, the acquired statokinesigrams are resampled at 25Hz using the SWARII algorithm [2]. Participants were

[1]The clinical trial (ID RCB 2014-A00222-45) was approved by Ethical Research Committees (CPP), Ile-de-France, University Paris VI. A written informed consent was obtained from all participants.

Algorithm 1 The proposed ts-AUC statistical test.

Input: X and Y are the points' coordinates of the trajectory (statokinesigram);
LS, OOB, M are vectors with the required hyper-parameters.
Output: AUC^*, RF^*, P^*, $p\text{-}value^*$.

Step 1: Exploration of the space of hyperparameters

for $i \in LS$ **do**
  for $j \in M$ **do**
    RF = RandForest (X, Y, LS_i, M_j)
    P = OOBpredict $(RF_{i,j})$
    U = Mann_Whitney_Utest_Statistic (P)
    $AUC_{i,j}$ = AUCestimation (U, Y)
  end for
end for

Step 2: Choose the best model and apply MWW

$(i^*, j^*) = \arg\max_{i \in LS, j \in M} AUC_{i,j}$
$AUC^* = AUC_{i^*,j^*}$
RF^* = RandForest $(X, Y, LS_{i^*}, M_{j^*})$
P^* = OOBpredict (RF^*)
$p\text{-}value^*$ = MWW (P^*, Y)

labeled as fallers (PS_F) if they had come to a lower level near the ground unintentionally at least once during the last six months [17]. Our analysis included only CoP trajectories' features that have been previously proposed by clinicians as indicators of postural impairment [14] (see Tab. 2 in Appendix for details).

Compared methods and settings. We compare the results obtained by the proposed ts-AUC with the Maximum Mean Discrepancy test (MMD) [10], which is a well-established multivariate test and state-of-the-art in terms of performance. We also compare with standard statistical testing approaches, which are usually employed in clinical studies. We check the p-values of all 17 features (i.e. $D = 17$) with the labels {'faller', 'non-faller'} using the non-parametric MWW test. Clinicians would typically report those features that were found statistically significant (e.g. with p-value $< \alpha = 0.05$) and any interesting non-significant finding.

In order to prevent the increase of the false positive probability, p-value adjustment procedures are applied. We use the Bonferroni correction, which is the most widely used p-value adjustment in biomedical research. Finally, we assess the effect of population size to the final result. We progressively reduce, uniformly at random, the number of PS_{NF} by a step of 10% (95% to 35%). All fallers were included, the test run 12 times and the percentages of significant results were compared (see Fig. 2).

3.1 Results

The ts-AUC and the MMD tests were applied to features derived from Eyes-Open and Eyes-Closed acquisitions separately. Both tests agreed that only the features derived by statokinesigrams of Eyes-Open significantly separated PS_F from PS_{NF}. Therefore, we will henceforth continue by presenting detailed analysis only for Eyes-Open features.

The most influential features were found to be the `VelocityY`, `VarianceY`, `AccelerationY`, `EllArea` (Confidence Ellipse area) and `MaxX` in descending order. Tab. 1 indicates those features that showed p-value < 0.05 and the decisions regarding statistical significance obtained after applying Bonferroni correction. In every row of Tab. 1, values at column 1 (p-values) were compared one by one to values at columns 3 of the same row (Bonferroni) and were found *always higher*. By these results, *none* of the features would reject the H_0 of two-sample MWW test.

Table 1: Significant and non-significant results of a univariate two-sample Mann-Whitney Wilcoxon (MWW) test, and the α level of significance before and after Bonferroni correction.

Feature	p-value MWW	α level before correction	α level after correction
`EllArea`	0.0045	0.05	0.0029
`VarianceY`	0.006	$\gg$	0.0029
`MaxY`	0.006	$\gg$	0.0029
`DistC`	0.007	$\gg$	0.0029
`RangeY`	0.008	$\gg$	0.0029
`VelocityY`	0.009	$\gg$	0.0029
`MaxX`	0.03	$\gg$	0.0029
`RangeX`	0.04	$\gg$	0.0029
`VarianceX`	0.04	$\gg$	0.0029
`MinY`	0.04	$\gg$	0.0029
`MinX`	>0.05	$\gg$	-
	-	-	
`VelocityX`	$\gg$	$\gg$	-
`AccX`	$\gg$	$\gg$	-
`F95X`	$\gg$	$\gg$	-
`AccY`	$\gg$	$\gg$	-
`F95Y`	$\gg$	$\gg$	-
`AngularDev`	$\gg$	$\gg$	

Each MWW p-value is compared horizontally with the corresponding α's.

Effect of population size. The population decrease through non-faller exclusion had an important effect to the performance of all tests. MMD and ts-AUC showed similar behavior. Specifically, the number of times that fallers/non-fallers were found statistically different was gradually decreasing. Multiple univariate testing showed most of the times that the groups could no be considered as statistically different.

4 DISCUSSION

The objective of this study was to introduce an easy, interpretable, and intuitive multivariate two-sample testing strategy for clinical research. It was shown that: a) the novel multivariate two-sample testing approach, ts-AUC, had equal performance with the state-of-the-art MMD test, with the additional element of providing feature importance assessment without further analysis, and b) the ts-AUC and the standard statistics (usually used in clinical studies), when both applied to the dataset of PS patients lead to contradictory conclusions. This disagreement seems to be linked to the relative conservatism of the traditional p-value correction strategies (increase of probability of false-negative findings) [8]. The medio-lateral movement has been reported as the most discriminative element between PS patients and age-matched controls [13] and seems that play a role in distinguishing fallers and non-fallers PS patients. However, the key-difference between fallers and non-fallers

PS was spotted in antero-posterior movement. Inreased antero-posterior movement was previously reported in PS patients while quiet standing with eyes open [11].

The use of OOB observations as cross-validation method has two basic advantages: a) provides faster results in the AUC maximization process, and b) allows the final MWW test to be applied once to the whole dataset, which is more intuitive for clinicians.

Concerning the unbiased feature importance, we believe that this addition is a cornerstone of the proposed approach and inline with the current clinicians' needs. While they need to know if two groups are (or are not) significantly separated, they are also interested to know the most influential features that lead to the reported result. Although the algorithm offers this convenience, we need to note that feature importance should be treated with extra care. The proposed approach tries to minimize the false conclusions concerning the importance of features when redundant features are present. According to [9], some of the collinear features (relevant to the phenomenon) will be in the final selection, and others will not. This issue is still under research and the current ts-AUC framework can integrate better solutions in the future. A general advice to clinicians can be to check for features exhibiting mutual information before the beginning of the testing process.

Interestingly, the gradual balancing (and decreasing of subset) between the groups of fallers and non-fallers, showed that the proposed test is less conservative than the multiple testing process (with correction). It would be fare to say that ts-AUC combines robustness while boosting the interpretability of the result. Exploratory studies, where a hypothesis about the structure of the dataset is not strictly defined in advance, could benefit from such multivariate approaches.

A methodological limitation of our study is that our dataset is slightly imbalanced, with many negative examples and few positive ones. In this case, metrics other than AUC (e.g. precision-recall (PR) curve, F1 score or area under the PR curve) could be more appropriate for avoiding possible overfitting [6]. We decided to keep the AUC criterion, as in [16], not only due to its theoretical association with the U-statistic, but also to fulfill one of our main objectives: to propose an algorithm as understandable, interpretable, and easy-to-implement as possible.

5 CONCLUSIONS

In this paper we showed that using the proposed ts-AUC two-sample test, which is a method oriented to clinical research, fallers and non-fallers patients who suffer from Parkinsonian syndromes (PS) can be distinguished by examining posturographic features that are derived following the basic Romberg protocol. This new approach was also able to reveal the posturographic features that are significantly different between the two groups (i.e. more discriminative). The separation appeared statistically less detectable when using traditional approaches such as multiple testing. Supplementary material about the algorithm, can be found at www.bargiotas.com/material. The results of our study highlighted that the ts-AUC, and other new multivariate methods based on machine learning, can play an important role in evaluating the usefulness of simple and inexpensive acquisition protocols as well as the extracted posturographic features.

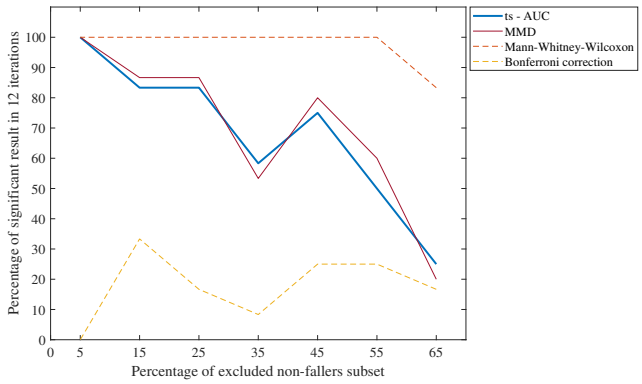

Figure 2: The average performance of two-sample testing approaches with smaller non-fallers population. ts-AUC and MMD have almost equal performance.

ACKNOWLEDGMENTS

The authors would like to thank Julien Audiffren for the implementation of the SWARII algorithm [2] used for statokinesigram pre-processing, and Albane Moreau for providing the additional database information concerning the PS patients. Argyris Kalogeratos was funded by the IdAML Chair hosted at ENS Paris-Saclay.

REFERENCES

[1] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. 2005. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research* 6, Apr (2005), 393–425.
[2] J. Audiffren and E. Contal. 2016. Preprocessing the Nintendo Wii board signal to derive more accurate descriptors of statokinesigrams. *Sensors* 16, 8 (2016), 1208.
[3] L. Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.
[4] S. Clémençon, G. Lugosi, and N. Vayatis. 2005. Ranking and scoring using empirical risk minimization. In *Int. Conf. on Comp. Learning Theory*. 1–15.
[5] C. Cortes and M. Mohri. 2004. AUC optimization vs. error rate minimization. In *Advances in Neural Information Processing Systems*. 313–320.
[6] J. Davis and M. Goadrich. 2006. The relationship between Precision-Recall and ROC curves. In *Int. Conf. on Machine Learning*. ACM, 233–240.
[7] J.A. Doornik and H. Hansen. 1996. *Out-of-bag estimation.* Technical Report. Dept. of Statistics, Univ. of California, Berkeley.
[8] R.J. Feise. 2002. Do multiple outcome measures require p-value adjustment? *BMC Medical Research Methodology* 2, 1 (2002), 8.
[9] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. 2010. Variable selection using random forests. *Pattern Recognition Letters* 31, 14 (2010), 2225–2236.
[10] A. Gretton, K.M. Borgwardt, M.J. Rasch, B. Schölkopf, and A. Smola. 2012. A kernel two-sample test. *Journal of Machine Learning Research* 13, Mar (2012), 723–773.
[11] G.K. Kerr, C.J. Worringham, M.H. Cole, P.F. Lacherez, J.M. Wood, and P.A. Silburn. 2010. Predictors of future falls in Parkinson disease. *Neurology* 75, 2 (2010), 116–124.
[12] J.M. Leach, M. Mancini, R.J. Peterka, T.L. Hayes, and F.B. Horak. 2014. Validating and calibrating the Nintendo Wii balance board to derive reliable center of pressure measures. *Sensors* 14, 10 (2014), 18244–18267.
[13] M. Mancini, P. Carlson-Kuhta, C. Zampieri, J.G. Nutt, L. Chiari, and F.B. Horak. 2012. Postural sway as a marker of progression in Parkinson's disease: a pilot longitudinal study. *Gait & Posture* 36, 3 (2012), 471–476.
[14] I. Melzer, N. Benjuya, and J. Kaplanski. 2004. Postural stability in the elderly: a comparison between fallers and non-fallers. *Age and Ageing* 33, 6 (2004), 602–607.
[15] M.S. Thiese, Z.C. Arnold, and S.D. Walker. 2015. The misuse and abuse of statistics in biomedical research. *Biochemia Medica* 25, 1 (2015), 5–11.
[16] N. Vayatis, M. Depecker, and S.J. Clémençcon. 2009. AUC optimization and the two-sample problem. In *Advances in Neural Information Processing Systems*. 360–368.
[17] A.A. Zecevic, A.W. Salmoni, M. Speechley, and A.A. Vandervoort. 2006. Defining a fall and reasons for falling: comparisons among the views of seniors, health care providers, and the research literature. *The Gerontologist* 46, 3 (2006), 367–376.