

# Ranking of Classification Algorithms in Terms of Mean–Standard Deviation Using A-TOPSIS

André G. C. Pacheco<sup>1</sup> · Renato A. Krohling<sup>1,2</sup>

Received: 22 October 2016 / Revised: 27 July 2017 / Accepted: 5 January 2018 /  
Published online: 13 January 2018  
© Springer-Verlag GmbH Germany, part of Springer Nature 2018

**Abstract** In classification problems when multiple algorithms are applied to different benchmarks a difficult issue arises, i.e., how can we rank the algorithms? In machine learning, it is common to run the algorithms several times and then a statistic is calculated in terms of means and standard deviations. In order to compare the performance of the algorithms, it is very common to employ statistical tests. However, these tests may also present limitations, since they consider only the means and not the standard deviations of the obtained results. In this paper, we present the so-called A-TOPSIS, based on Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), to solve the problem of ranking and comparing classification algorithms in terms of means and standard deviations. We use two case studies to illustrate the A-TOPSIS for ranking classification algorithms and the results show the suitability of A-TOPSIS to rank the algorithms. The presented approach can be applied to compare the performance of stochastic algorithms in machine learning. Lastly, to encourage researchers to use the A-TOPSIS for ranking algorithms, we also presented in this work an easy-to-use A-TOPSIS web framework.

**Keywords** Ranking algorithms · Machine learning · Classification algorithms · Comparison of algorithms · TOPSIS · Statistical tests

---

✉ André G. C. Pacheco  
pacheco.comp@gmail.com  
Renato A. Krohling  
krohling.renato@gmail.com

- <sup>1</sup> Graduate Program in Computer Science, PPGI, UFES - Federal University of Espírito Santo, Av. Fernando Ferrari 514, Vitória, ES CEP 29060-270, Brazil
- <sup>2</sup> Production Engineering Department, UFES - Federal University of Espírito Santo, Av. Fernando Ferrari 514, Vitória, ES CEP 29060-270, Brazil

## 1 Introduction

In machine learning, and more precisely in classification problems, it is very common applying different algorithms to many benchmarks several times. Normally, the performance of the algorithms is analyzed using the mean and the standard deviation of some known metric, such as the classification accuracy. Next, we need to compare the algorithms and a difficult question arises: how to compare these algorithms effectively? The first answer to this question is to use statistical tests, i.e., parametric and/or nonparametric. The statistical tests can detect if there are differences between the performances of the algorithms [4,5]. One problem is if there are differences, which algorithm is the best, the second best, and the worst? Using nonparametric statistical tests, it is necessary to make pairwise and multiple comparisons among the algorithms. Obviously, the number of tests required increases greatly with the number of algorithms being analyzed. This is problematic, firstly because of the tiresome work of comparing each pair of algorithms; secondly, and more importantly, the probability of making a mistake increases. In addition, these tests may also present limitations, since they consider only the means and not the standard deviations of the obtained results.

Over the past few years, some approaches have been proposed in order to rank classification algorithms. Brazdil and Soares [2] presented three methods to generate rankings of classification algorithms. However, these methods are not robust and sometimes their results do not match with the statistical tests. Peng et al. [13] developed a decision-making framework to rank classification algorithms. Nonetheless, this framework does not consider the standard deviation of the algorithms' performance. Moreover, the authors do not compare their methods with the statistical tests. Kotthoff [10] investigated ranking approaches to select the most appropriate algorithm for solving a particular problem. In this case, his goal is to tackle the Algorithm Selection Problem [14], which is slightly different from ranking algorithms based on performance in different benchmarks.

Recently, Krohling et al. [9] presented a new approach to support the selection of the best algorithms by using the Hellinger distance [11]. This approach, called Hellinger-TOPSIS, provides a rank order of the algorithms in an easy and direct way, using the mean and the standard deviation of the performance of the algorithms. However, the Hellinger-TOPSIS presents some shortcomings. Firstly, the mean and the standard deviation of the algorithms' performance have the same importance. Usually, the mean of the performance is more important than the standard deviation. In the Hellinger-TOPSIS we cannot control the influence of these two parameters. Second, if any algorithm in the group is deterministic, i.e, the results obtained are described just by the means, and it must be compared with other stochastic ones, the Hellinger-TOPSIS cannot handle such a case, because in the algorithm the standard deviation must be different from zero.

In our previous work, we proposed the A-TOPSIS [8], a new approach that provides a rank order of the evolutionary algorithms in cases where the performance of the algorithms are expressed in terms of means and standard deviations. In this work, we extend our previous approach by providing an in-depth investigation for two case studies for classification problems. In addition, we develop an easy-to-use web

framework for A-TOPSIS. The remainder of this paper is organized as follows: Sect. 2 presents a background in decision-making and in the TOPSIS. In Sect. 3, we present the approach based on TOPSIS to deal with data matrix consisting of the performance of algorithms in terms of means and standard deviations and we briefly describe the web framework developed. In Sect. 4, we present simulation results for two case studies involving the classification task in order to illustrate the suitability of the presented approach. In Sect. 5, conclusions and directions for future work are given.

## 2 Background in Decision-Making and TOPSIS

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS), developed by Hwang and Yoon [6], is a technique to evaluate the performance of alternatives through the similarity with the ideal solution. According to this technique, the best alternative would be one that is closest to the positive-ideal solution and farthest from the negative-ideal solution. The positive-ideal solution is the one that maximizes the benefit criteria and minimizes the cost criteria. The negative-ideal solution maximizes the cost criteria and minimizes the benefit criteria. In summary, the positive-ideal solution is composed of all the best values attainable for the criteria, and the negative-ideal solution consists of all the worst values attainable for the criteria. The interested reader shall refer to Behzadian et al. [1] for a broad survey about TOPSIS.

Let us consider the decision matrix  $A$ , which consists of *alternatives* and *criteria*, described by:

$$A = \begin{matrix} & C_1 & \dots & C_n \\ \begin{matrix} A_1 \\ \dots \\ A_m \end{matrix} & \begin{pmatrix} x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} \end{matrix} \tag{1}$$

where  $A_1, A_2, \dots, A_m$  are viable alternatives, and  $C_1, C_2, \dots, C_n$  are criteria,  $x_{ij}$  indicates the rating of the alternative  $A_i$  with respect to criterion  $C_j$ . The weight vector  $W = (w_1, w_2, \dots, w_n)$  is composed of the individual weights  $w_j (j = 1, \dots, n)$ , for each criterion  $C_j$  and satisfies  $\sum_{j=1}^n w_j = 1$ . In general, the criteria can be classified into two types: *benefit* and *cost*. The *benefit* criterion means that a higher value is better, while for the *cost* criterion the opposite is valid. The data of the decision matrix  $A$  come from different sources, so it is necessary to normalize it in order to transform it into a dimensionless matrix, which allows the comparison of the various criteria. In this work, we use the normalized decision matrix  $R = [r_{ij}]_{m \times n}$  with  $i = 1, \dots, m$ , and  $j = 1, \dots, n$ . The normalized value  $r_{ij}$  is calculated as:

$$r_{ij} = \frac{x_{ij}}{\sqrt{\sum_{i=1}^m x_{ij}^2}}, \text{ with } i = 1, \dots, m; j = 1, \dots, n. \tag{2}$$

or

$$r_{ij} = \frac{x_{ij}}{x_{i \max}}, \text{ with } i = 1, \dots, m; j = 1, \dots, n. \tag{3}$$

The normalized decision matrix  $R$  represents the relative rating of the alternatives. After normalization, one calculates the weighted normalized decision matrix  $P = [p_{ij}]_{m \times n}$  with  $i = 1, \dots, m$ , and  $j = 1, \dots, n$  by multiplying the normalized decision matrix by its associated weights. The weighted normalized value  $p_{ij}$  is calculated as:

$$p_{ij} = w_j \cdot r_{ij} \text{ with } i = 1, \dots, m, \text{ and } j = 1, \dots, n. \tag{4}$$

The TOPSIS is described in the following steps [6, 7]:

*Step 1* Identify the positive ideal solutions  $A^+$  (benefits) and negative ideal solutions  $A^-$  (costs) as follows:

$$A^+ = (p_1^+, p_2^+, \dots, p_n^+) \tag{5}$$

$$A^- = (p_1^-, p_2^-, \dots, p_n^-) \tag{6}$$

where  $p_j^+ = (\max_i p_{ij}, j \in J_1; \min_i p_{ij}, j \in J_2)$  and  $p_j^- = (\min_i p_{ij}, j \in J_1; \max_i p_{ij}, j \in J_2)$ ;  $J_1$  and  $J_2$  represent the criteria *benefit* and *cost*, respectively.

*Step 2* Calculate the Euclidean distances from the positive ideal solution  $A^+$  (benefits) and the negative ideal solution  $A^-$  (costs) of each alternative  $A_i$ , respectively as follows:

$$d_i^+ = \sqrt{\sum_{j=1}^n (d_{ij}^+)^2} \tag{7}$$

$$d_i^- = \sqrt{\sum_{j=1}^n (d_{ij}^-)^2} \tag{8}$$

where  $d_{ij}^+ = p_j^+ - p_{ij}$ , with  $i = 1, \dots, m$  and  $d_{ij}^- = p_j^- - p_{ij}$ , with  $i = 1, \dots, m$ .

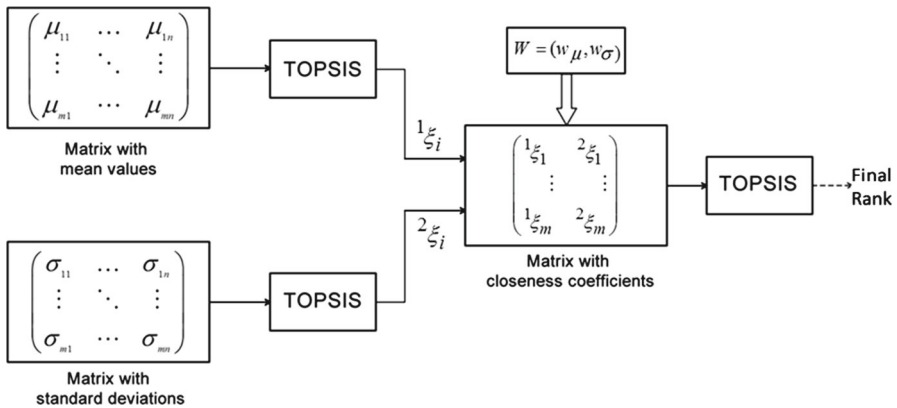
*Step 3* Calculate the relative closeness coefficients  $\xi_i$  for each alternative  $A_i$  with respect to the positive ideal solution as given by:

$$\xi_i = \frac{d_i^-}{d_i^+ + d_i^-} \tag{9}$$

*Step 4* Rank the alternatives according to the relative closeness. The best alternatives are those that have higher value  $\xi_i$  and therefore should be chosen.

### 3 A-TOPSIS: An Approach Based on TOPSIS for Ranking Algorithms

The A-TOPSIS is an approach for ranking algorithms that uses the TOPSIS as a building block. Its main idea is to rank a group of algorithms by using the means and the standard deviations of their performance. Let us consider a group of  $m$  algorithms



**Fig. 1** Illustration of the A-TOPSIS: an approach for ranking algorithms in terms of mean and standard deviations

performed to  $n$  benchmarks. We can set a decision matrix using all the performance means and the standard deviations of each algorithm for each benchmark as follows:

$$D = \begin{matrix} A_1 \\ \vdots \\ A_m \end{matrix} \begin{pmatrix} C_1 & \dots & C_n \\ x_{11} & \dots & x_{1n} \\ \vdots & \ddots & \vdots \\ x_{m1} & \dots & x_{mn} \end{pmatrix} = \begin{pmatrix} (\mu_{11}, \sigma_{11}) & \dots & (\mu_{1n}, \sigma_{1n}) \\ \vdots & \ddots & \vdots \\ (\mu_{m1}, \sigma_{m1}) & \dots & (\mu_{mn}, \sigma_{mn}) \end{pmatrix}$$

where  $A_1, A_2, \dots, A_m$  are alternatives,  $C_1, C_2, \dots, C_n$  are criteria,  $x_{ij}$  indicates the rating of the alternative  $A_i$  with respect to criterion  $C_j$  described in terms of its mean and standard deviations  $(\mu_{ij}, \sigma_{ij})$ , respectively. As we can note, for the A-TOPSIS, the alternatives consist of the algorithms and the criteria are the benchmark problems.

The decision matrix  $D$  can be split into two matrices, given by  $D = \{M_\mu, M_\sigma\}$ :

$$M_\mu = \begin{pmatrix} \mu_{11} & \dots & \mu_{1n} \\ \vdots & \ddots & \vdots \\ \mu_{m1} & \dots & \mu_{mn} \end{pmatrix} \quad M_\sigma = \begin{pmatrix} \sigma_{11} & \dots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \dots & \sigma_{mn} \end{pmatrix}$$

where  $M_\mu$  and  $M_\sigma$  are the matrices of the means and the standard deviations of the algorithms performance for each benchmark problem. From these matrices, we developed a new framework combining the TOPSIS as illustrated in Fig. 1.

### 3.1 A-TOPSIS Algorithm

Next, we present the step-by-step of the proposed framework:

*Step 1* Normalize the matrices  $M_\mu$  and  $M_\sigma$ .

Step 2 Identify the positive ideal solutions  $A^+$  (benefits) and negative ideal solutions  $A^-$  (costs) for each matrix as follows:

$$A^+ = (p_1^+, p_2^+, \dots, p_n^+) \tag{10}$$

$$A^- = (p_1^-, p_2^-, \dots, p_n^-) \tag{11}$$

where  $p_j^+ = (\max_i p_{ij}, j \in J_1; \min_i p_{ij}, j \in J_2)$  and  $p_j^- = (\min_i p_{ij}, j \in J_1; \max_i p_{ij}, j \in J_2)$ ;  $J_1$  and  $J_2$  represent the criteria *benefit* and *cost*, respectively.

Step 3 Calculate the Euclidean distances from the positive ideal solution  $A^+$  (benefits) and the negative ideal solution  $A^-$  of each alternative  $A_i$ , respectively as follows:

$$d_i^+ = \sqrt{\sum_{j=1}^n (p_j^+ - p_{ij})^2} \text{ with } i = 1, \dots, m. \tag{12}$$

$$d_i^- = \sqrt{\sum_{j=1}^n (p_j^- - p_{ij})^2} \text{ with } i = 1, \dots, m. \tag{13}$$

Step 4 Calculate the relative closeness coefficients for each alternative  $\xi_i$  with respect to positive ideal solution as:

$$\xi_i = \frac{d_i^-}{d_i^+ + d_i^-} \text{ with } i = 1, \dots, m. \tag{14}$$

Step 5 After calculating the vector  $\xi_i$  for both decision matrices, we obtain a data matrix that is made up of the two vectors of the relative closeness coefficients, as given by:

$$C = \begin{pmatrix} {}^1\xi_1 & {}^2\xi_1 \\ \vdots & \vdots \\ {}^1\xi_m & {}^2\xi_m \end{pmatrix} \tag{15}$$

In this case, to each of the vectors, it is assigned a weight  $W = (w_1, w_2) = (w_\mu, w_\sigma)$ , where  $w_\mu$  and  $w_\sigma$  represent the weight assigned to the criteria means, and standard deviations, respectively, which satisfies  $w_\mu + w_\sigma = 1$ . One can now obtain the weighted relative-closeness coefficients matrix by introducing the importance weights to each one of the relative-closeness coefficient vector, as given by:

$$C = \begin{pmatrix} w_1 {}^1\xi_1 & w_2 {}^2\xi_1 \\ \vdots & \vdots \\ w_1 {}^1\xi_m & w_2 {}^2\xi_m \end{pmatrix} \tag{16}$$

From this stage on, the method continues by applying the standard TOPSIS to the resulting matrix in order to identify the global ranking.

*Step 6* Identify the global positive ideal solution  $A_G^+$  and the global negative ideal solution  $A_G^-$ , respectively, as follows:

$$A_G^+ = (p_{G1}^+, p_{G2}^+) = \left( \max_i {}^l \xi_i, \quad l \in J_1; \min_i {}^l \xi_i \quad l \in J_2 \right). \tag{17}$$

$$A_G^- = (p_{G1}^-, p_{G2}^-) = \left( \min_i {}^l \xi_i, \quad l \in J_1; \max_i {}^l \xi_i \quad l \in J_2 \right). \tag{18}$$

where  $J_1$  and  $J_2$  represent the criteria *benefit* and *cost*, respectively.

*Step 7* Calculate to each alternative  $A_i$  the distances from the global positive ideal solution  $A_G^+$  and from the global negative ideal solution  $A_G^-$ , respectively, as follows:

$$d_{Gi}^+ = \sqrt{\sum_{l=1}^2 \left( {}^l \xi_i - p_{G_l}^+ \right)^2} \text{ with } i = 1, \dots, m. \tag{19}$$

$$d_{Gi}^- = \sqrt{\sum_{l=1}^2 \left( {}^l \xi_i - p_{G_l}^- \right)^2} \text{ with } i = 1, \dots, m. \tag{20}$$

*Step 8* Calculate the global relative-closeness coefficients  $\xi_{Gi}$  for each alternative  $A_i$  with respect to global positive ideal solution  $A_G^+$  as:

$$\xi_{Gi} = \frac{d_{Gi}^-}{d_{Gi}^- + d_{Gi}^+} \tag{21}$$

*Step 9* Rank the alternatives according to the relative closeness coefficients. The best alternatives are those that have higher value  $\xi_{Gi}$  and therefore should be chosen.

### 3.2 A-TOPSIS Web Framework

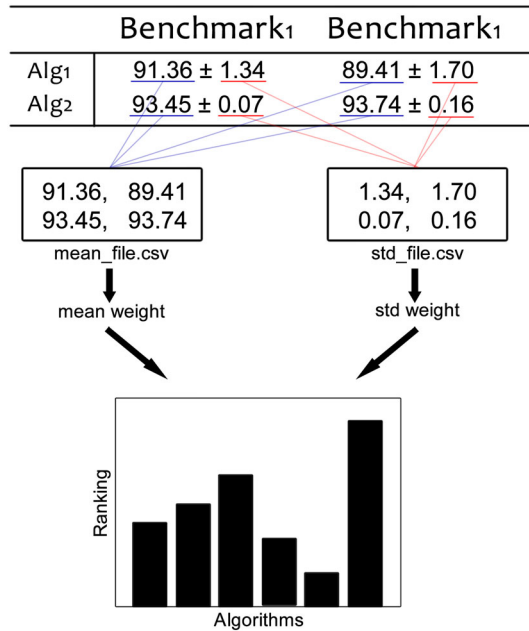
In order to encourage researchers and practitioners from different areas of knowledge to use the A-TOPSIS for ranking algorithms, we provide an easy-to-use web framework. As shown in Fig. 2, to use this framework the user needs to set the matrices  $M_\mu$  and  $M_\sigma$  as .csv files and the value of the weights for each one. Thereby, the framework provides the graph bar rank and the values of the closeness coefficients.

The A-TOPSIS framework can be easily used by accessing the web address <http://www.inf.ufes.br/~agcpacheco/alg-ranking/>.

## 4 Simulation Results

In this section, we present two case studies involving classification problems. In order to compare our results, we also apply the Hellinger-TOPSIS for each case. As the

**Fig. 2** The A-TOPSIS framework



Hellinger-TOPSIS cannot handle a standard deviation equal to zero, we set a very small value as the standard deviation in cases where this occurs. Lastly, we used the nonparametric Friedman test followed by Wilcoxon test as a pos hoc, both with  $p_{value} = 0.05$ , in order to certify the quality of the rank. For more details about the statistical tests performed in this section, the reader may refer to Derrac et al. [4].

### 4.1 Case Study I

In this case study, we have an ensemble of classifiers, containing four classifiers: feedforward neural network (FNN), extreme learning machine (ELM), discriminative restricted Boltzmann machine (DRBM) and K-nearest neighbors (KNN). In addition, we have three aggregation methodologies: the average of the supports (AVG), the majority voting (MV) and the Choquet integral (CHO) [12]. All these classifiers were applied to 12 benchmarks, and their performance for each benchmark is described in Table 1. Our goal is to rank the seven algorithms according to their performance. Therefore, the decision matrix  $D = \{M_\mu, M_\sigma\}$ , presented in Sect. 3, is set by using the values described in Table 1.

As we can see in Table 1, the KNN algorithm does not have a standard deviation because it was used with just one value of  $k$ . Therefore, we divided this case study into two parts. First, we remove the KNN and consider the remaining classifiers. Second, we consider all seven classifiers setting the KNN standard deviation equal to zero. We decided to do it to show the ranking differences between the A-TOPSIS and the Hellinger-TOPSIS when we include an algorithm with standard deviation equal to

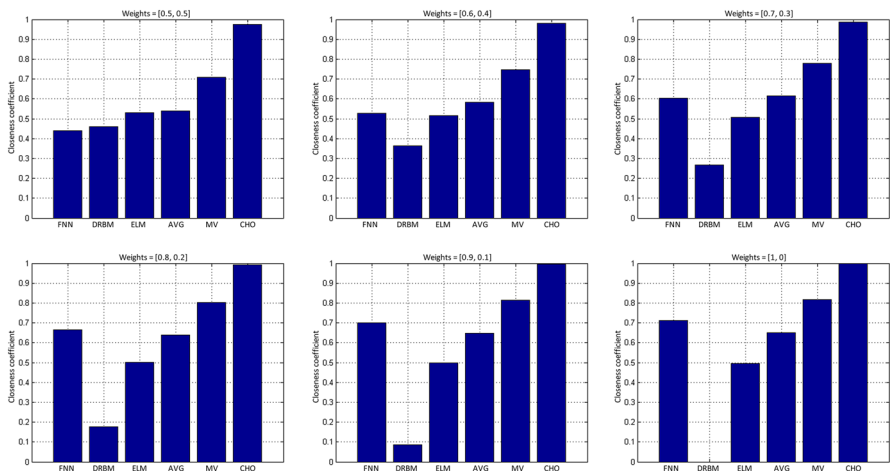


**Table 1** The classifiers performance for each benchmark in terms of mean and standard deviation of the classification accuracy—case study I

Classifiers	Benchmarks											
	Susy	Higgs	Covtype	DNA	Isolet	Cancer	Cred. Aus	Diabetic	Iris	Spam	Statlog	Wine
FNN	78.14 ± 0.65	63.21 ± 1.19	75.22 ± 1.09	91.36 ± 1.34	89.41 ± 1.7	94.87 ± 0.62	83.05 ± 1.07	71.02 ± 2.31	95.62 ± 1.36	93.61 ± 0.8	99.59 ± 0.06	95.91 ± 2.16
	DRBM	76.39 ± 0.32	63.4 ± 0.3	66.25 ± 0.17	93.45 ± 0.07	93.74 ± 0.16	95.39 ± 0.28	82.88 ± 0.8	60.48 ± 1.69	89.4 ± 0.95	90.8 ± 0.56	92.08 ± 0.04
ELM	79.39 ± 0.29	63.99 ± 0.09	76.01 ± 0.11	90.59 ± 0.75	86.81 ± 0.6	95.07 ± 0.5	83.38 ± 0.81	72.37 ± 1.09	94.81 ± 1.96	89.43 ± 0.57	98.3 ± 0.08	91.13 ± 2.8
	KNN	70.88 ± 0	59.84 ± 0	75.81 ± 0	85.98 ± 0	88.24 ± 0	67.63 ± 0	61.73 ± 0	95.55 ± 0	72.75 ± 0	98.73 ± 0	67.92 ± 0
AVG	78.38 ± 0.59	64.01 ± 0.96	71.84 ± 0.42	92.18 ± 0.92	93.73 ± 0.26	95.38 ± 0.35	83.14 ± 0.55	68.84 ± 3.03	95.11 ± 1.47	93.4 ± 0.61	98.73 ± 0.01	96.16 ± 1.44
	MV	78.14 ± 0.54	63.75 ± 0.64	75.75 ± 0.24	92.64 ± 0.57	93.93 ± 0.31	83.15 ± 0.98	70.05 ± 2.24	95.62 ± 1.42	92.02 ± 0.37	99.32 ± 0.04	95.22 ± 1.76
CHO	78.58 ± 0.51	64.7 ± 0.72	76.85 ± 0.29	93.69 ± 0.38	93.75 ± 0.16	95.57 ± 0.35	83.52 ± 0.85	71.42 ± 0.97	95.7 ± 1.29	93.78 ± 0.65	98.73 ± 0.01	97.18 ± 1.34

**Table 2** Rank by varying the values of the weight—case study I, part I

Weight variation [mean, std]	Ranking
[0.5, 0.5]	CHO > MV > AVG > ELM > DRBM > FNN
[0.6, 0.4]	CHO > MV > AVG > FNN > ELM > DRBM
[0.7, 0.3]	CHO > MV > AVG > FNN > ELM > DRBM
[0.8, 0.2]	CHO > MV > AVG > FNN > ELM > DRBM
[0.1, 0.9]	CHO > MV > AVG > FNN > ELM > DRBM
[1, 0]	CHO > MV > AVG > FNN > ELM > DRBM



**Fig. 3** Rank in bar graph for each weights configuration—case study I, part I

zero (recall the Hellinger-TOPSIS considers the mean and the standard deviation with the same importance). For both parts, we carry out a sensitivity study by varying the weights for the mean and the standard deviation, respectively.

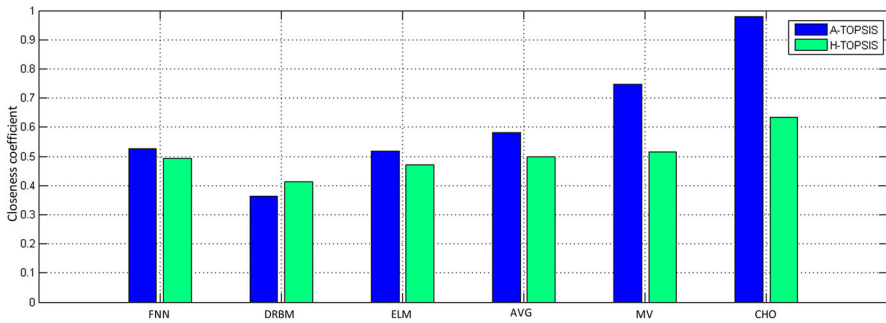
4.1.1 Case Study I: Part I

From the Table 1, we remove the KNN and maintain the remaining algorithms. Thereby, the decision matrix for this experiment has six alternatives (algorithms) and 12 criteria (benchmarks). In Table 2 is described the rank provided by the A-TOPSIS by varying the weights for the mean and the standard deviation.

As we can see, the first, second and third place in the rank do not change regardless the weight. In fact, the only change in the rank occurs when the values of the weights become [0.6, 0.4]. In this case, the FNN rises to the fourth place, the ELM goes down to the fiftieth place and the DRBM goes to the last place. Varying the weights from [0.6, 0.4] to [1, 0] does not change the rank. In Fig. 3 is illustrated the raking in the bar graph for each weights configuration.

**Table 3** Rank comparison between A-TOPSIS and H-TOPSIS—part I

Method	Ranking
A-TOPSIS	CHO > MV > AVG > FNN > ELM > DRBM
Hellinger-TOPSIS	CHO > MV > AVG > FNN > ELM > DRBM



**Fig. 4** Rank in bar graph for A-TOPSIS and H-TOPSIS—case study I, part I

**Table 4** Wilcoxon test pairwise comparison with  $p$  less than 0.05—case study I, part I

Pairwise	$p$	Pairwise	$p$
FNN–CHO	0.009277	ELM–CHO	0.042480
DRBM–AVG	0.015137	AVG–CHO	0.000977
DRBM–MV	0.026855	MV–CHO	0.009277
DRBM–CHO	0.000488	–	–

We compare the results obtained by A-TOPSIS with the Hellinger-TOPSIS. Since the rank provided by the A-TOPSIS becomes stable with weights equal to [0.6, 0.4], we chose these values for this comparison. In Table 3 is presented the rank for each methodology, which is also depicted in Fig. 4 in the bar graph. According to the presented results, both methods obtained the same rank.

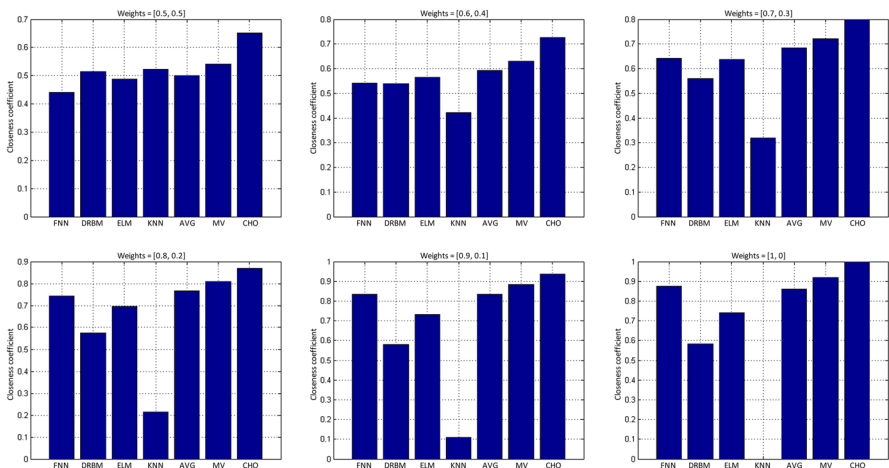
The Friedman test for this experiment provides  $p_{value} = 0.00005$ , leading to reject  $H_0$ . Then, we perform the pairwise comparisons using the Wilcoxon test. According to the results presented in Table 4, the CHO classifier is significantly different when compared to the other ones. Furthermore, the DRBM classifier is significantly different than AVG, MV and CHO. Thus, the statistical tests indicate that the CHO classifier is the best algorithm and the DRBM is the worst one. This finding is consistent with the results obtained by A-TOPSIS. Nonetheless, the statistical tests cannot provide a rank with all the classifiers as A-TOPSIS does.

4.1.2 Case Study I: Part II

In this experiment, we consider all the algorithms in Table 1. Thereby, the decision matrix for this experiment has seven alternatives (algorithms) and 12 criteria (bench-

**Table 5** Rank by varying the values of the weights—case study I, part II

Weight variation [mean, std]	Ranking
[0.5, 0.5]	CHO > MV > KNN > DRBM > AVG > ELM > FNN
[0.6, 0.4]	CHO > MV > AVG > ELM > FNN > DRBM > KNN
[0.7, 0.3]	CHO > MV > AVG > FNN > ELM > DRBM > KNN
[0.8, 0.2]	CHO > MV > AVG > FNN > ELM > DRBM > KNN
[0.1, 0.9]	CHO > MV > AVG > FNN > ELM > DRBM > KNN
[1, 0]	CHO > MV > AVG > FNN > ELM > DRBM > KNN



**Fig. 5** Rank in bar graph for each values of the weight—case study I, part II

marks). In Table 5 is described the rank provided by the A-TOPSIS by varying the weights for the mean and the standard deviation.

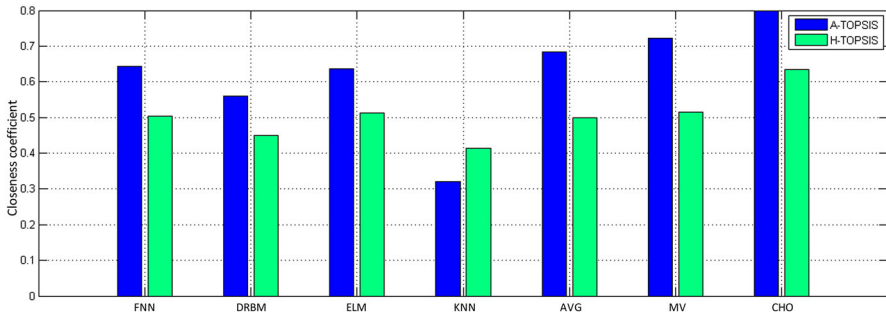
As we can notice in Table 5, for all weights the first and the second place in the rank do not change. For the weights equal to [0.5, 0.5], the KNN reaches the third place in the rank. However, when the weights are set equal to [0.6, 0.4], only 10% of variation, the KNN goes down to the last place. Moreover, for these weights, the ELM rises from the sixth to the fourth place. Nevertheless, when the weights are equal to [0.7, 0.3], the ELM and the FNN switch their positions. From the weights [0.7, 0.3] to [1, 0], the rank become stable and does not change anymore. In Fig. 5 is illustrated the rank in bar graph for each weights configuration.

Again, we compare the results obtained by A-TOPSIS with the Hellinger-TOPSIS. For this experiment, we choose the stable weights [0.7, 0.3]. In Table 6 is presented the rank for each methodology, which is also depicted in Fig. 6 in the bar graph.

In Table 6, we can easily check that the ranking of the alternatives CHO, FNN, DRBM and KNN are the same in both methods. Thus, the ranking of the best and worst alternatives are kept. On the other hand, the ranking of the alternatives MV, AVG and

**Table 6** Rank comparison between A-TOPSIS and H-TOPSIS—case study I, part II

Method	Ranking
A-TOPSIS	CHO > MV > AVG > FNN > ELM > DRBM > KNN
Hellinger-TOPSIS	CHO > ELM > MV > FNN > AVG > DRBM > KNN



**Fig. 6** Rank in bar graph for A-TOPSIS and H-TOPSIS—case study I, part II

**Table 7** Wilcoxon test pairwise comparison with  $p$  less than 0.05—case study I, part II

Pairwise	$p$	Pairwise	$p$
FNN–KNN	0.004883	ELM–CHO	0.042480
FNN–CHO	0.009277	KNN–AVG	0.009766
DRBM–AVG	0.015137	KNN–MV	0.003418
DRBM–MV	0.026855	KNN–CHO	0.000977
DRBM–CHO	0.000488	AVG–CHO	0.000977
ELM–KNN	0.042480	MV–CHO	0.009277

ELM have changed their positions. Comparing the rank in both experiments (Tables 3, 6), we observe that the A-TOPSIS include the KNN in the last position and maintain the ranking for the remaining algorithms. Conversely, the Hellinger-TOPSIS does not do the same. Therefore, we conclude that the inclusion of the KNN directly affects in the Hellinger-TOPSIS ranking. This happened because the Hellinger-TOPSIS does not allow us to control the influence of the mean and the standard deviation in its algorithm.

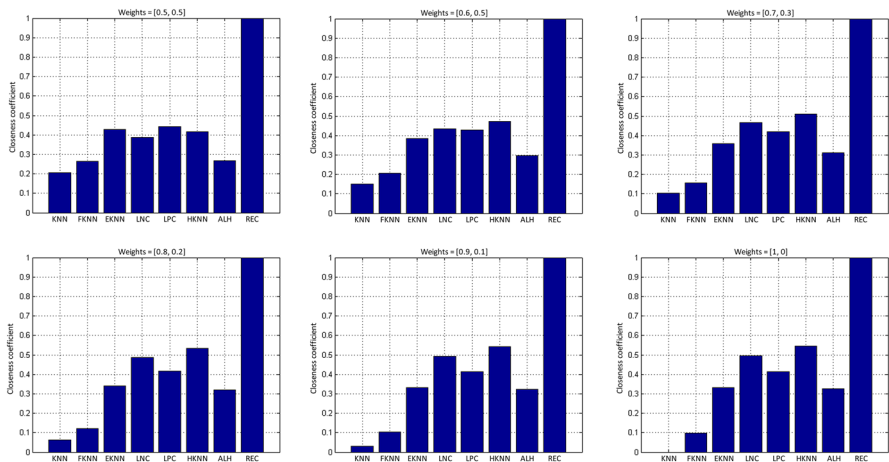
Similarly to the previous experiment, the Friedman test for this experiment provides  $pvalue = 0.00007$ , leading to reject  $H_0$ . Next, we perform the pairwise comparisons using the Wilcoxon test. According to the results presented in Table 7, the CHO classifier is significantly different when compared to the other ones. In addition, the KNN is also significantly different than the others, except DRBM. Lastly, the DRBM is significantly different comparing to AVG, MV and CHO. Thus, the statistical tests indicate that the CHO classifier is the best algorithm and the KNN and the DRBM are the worst ones. Also in this case, this finding is consistent with the results obtained by A-TOPSIS.

**Table 8** The classifiers performance for each benchmark in terms of mean and standard deviation of the classification accuracy—case study 2

Classifiers	Benchmarks									
	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10
KNN	3.49 ± 0.49	3.35 ± 0.39	25.83 ± 0.70	19.42 ± 0.79	30.82 ± 1.40	14.10 ± 1.60	4.40 ± 0.60	3.94 ± 0.37	18.96 ± 1.71	29.84 ± 0.55
FKNN	3.49 ± 0.24	3.13 ± 0.30	26.28 ± 0.61	15.96 ± 0.68	30.45 ± 1.82	14.10 ± 1.04	4.53 ± 0.56	2.40 ± 0.36	19.04 ± 0.93	30.50 ± 1.03
EKNN	2.26 ± 0.56	2.96 ± 0.57	25.76 ± 1.01	10.77 ± 1.25	30.71 ± 0.51	14.10 ± 1.24	5.07 ± 0.37	3.90 ± 0.32	19.19 ± 1.21	31.16 ± 1.40
LMC	2.70 ± 0.45	2.35 ± 0.47	24.85 ± 0.86	11.11 ± 1.05	31.59 ± 1.84	11.90 ± 1.47	4.27 ± 0.76	2.00 ± 0.24	19.41 ± 1.90	26.84 ± 1.08
LPC	2.60 ± 0.49	2.68 ± 0.58	25.03 ± 0.68	12.36 ± 0.59	32.0 ± 1.83	11.90 ± 1.21	4.00 ± 0.82	2.10 ± 0.31	19.04 ± 1.0	27.10 ± 0.40
HKNN	2.04 ± 0.84	2.57 ± 0.73	25.31 ± 0.83	11.44 ± 0.55	29.63 ± 2.34	11.81 ± 1.66	4.00 ± 0.82	1.98 ± 0.49	20.44 ± 1.29	24.64 ± 1.28
ALH	2.48 ± 0.51	3.07 ± 0.44	29.92 ± 0.88	11.62 ± 0.79	31.37 ± 2.36	10.86 ± 1.28	4.67 ± 0.94	2.00 ± 0.32	22.30 ± 1.27	26.72 ± 1.20
REC	1.58 ± 0.46	2.13 ± 0.32	24.35 ± 0.75	6.10 ± 0.43	28.48 ± 1.20	10.76 ± 0.93	4.00 ± 0.00	0.79 ± 0.13	18.52 ± 0.74	24.40 ± 0.90

**Table 9** Rank by varying the values of the weights—case study II

Weight variation [mean, std]	Ranking
[0.5, 0.5]	REC > LPC > EKNN > HKNN > LNC > ALH > FKNN > KNN
[0.6, 0.4]	REC > HKNN > LNC > LPC > EKNN > ALH > FKNN > KNN
[0.7, 0.3]	REC > HKNN > LNC > LPC > EKNN > ALH > FKNN > KNN
[0.8, 0.2]	REC > HKNN > LNC > LPC > EKNN > ALH > FKNN > KNN
[0.9, 0.1]	REC > HKNN > LNC > LPC > EKNN > ALH > FKNN > KNN
[1, 0]	REC > HKNN > LNC > LPC > ALH > EKNN > FKNN > KNN



**Fig. 7** Rank in bar graph for each values of the weight—case study II

### 4.2 Case Study II

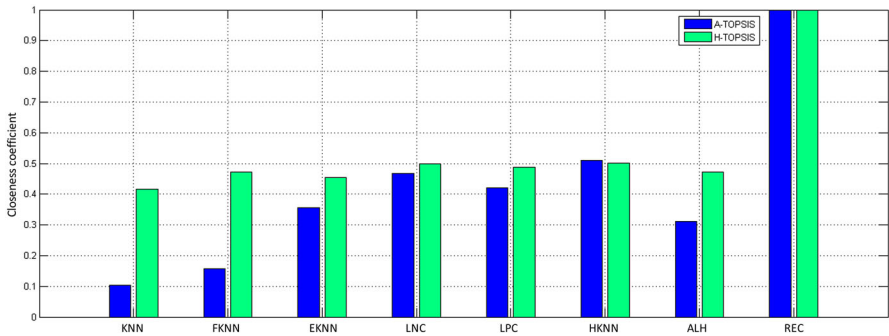
This case study presented by Wen et al. [15] consists in a classification problem with eight classifiers applied to 10 benchmarks. In Table 8 is described the performance of the classifiers for each benchmark. Similar to the case study 1, our goal is to find the rank of the classifiers according to their performance. It is worth mentioning that in this case study the authors used the error rate as accuracy. The A-TOPSIS can easily handle with this just changing the criterion from benefit to cost, i.e., the smaller the value is, the better.

As in this case study we have eight algorithm and ten benchmarks, the decision matrix has eight alternatives and ten criteria. In Table 9 is described the rank provided by the A-TOPSIS by varying the weights for the mean and the standard deviation.

As we can notice in Table 9, the first and the last place are the same for all weights. When the weights are varied from [0.5, 0.5] to [0.6, 0.4] the classifiers HKNN, LNC, LPC and EKNN switch their positions. Varying the weights from [0.6, 0.4] to [0.9, 0.1] the rank does not change. Lastly, when the weights become [1, 0], the classifiers

**Table 10** Rank comparison between A-TOPSIS and H-TOPSIS—case study II

Method	Ranking
A-TOPSIS	REC > HKNN > LNC > LPC > EKNN > ALH > FKNN > KNN
Hellinger-TOPSIS	REC > LNC ≈ HKNN > LPC > ALH > FKNN > EKNN > KNN



**Fig. 8** Rank in bar graph for A-TOPSIS and H-TOPSIS—case study II

**Table 11** Wilcoxon test pairwise comparison with  $p$  less than 0.05—case study II

Pairwise	$p$	Pairwise	$p$
KNN–LMC	0.019531	EKNN–REC	0.001953
KNN–LPC	0.037109	LMC–REC	0.001953
KNN–HKNN	0.027344	LPC–REC	0.003906
KNN–REC	0.001953	HKNN–ALH	0.027344
FKNN–LMC	0.048828	HKNN–REC	0.003906
FKNN–HKNN	0.027344	ALH–REC	0.001953
FKNN–REC	0.001953	–	–

EKNN and ALH switch their positions. In Fig. 7 is illustrated the rank in bar graph for each weights configuration.

In Table 10 is described the rank provided by the A-TOPSIS and Hellinger-TOPSIS. For the A-TOPSIS we chose the stable weights [0.7, 0.3]. In addition, in Fig. 8 is depicted the rank of each approach in bar graph.

From Table 10, we observe that there are some differences in the ranks. For both ranks, the first and the last places do not change. However, in the Hellinger-TOPSIS rank, the position of the classifiers LNC and HKNN are too close, therefore, they are tied in second place. Furthermore, as we can see in Fig. 8, even though we can distinguish the ranking for the classifiers FKNN, EKNN, LPC and ALH in Hellinger-TOPSIS rank, the values of the closeness coefficients are too close. This issue does not occur in A-TOPSIS.

For the case study II, the Friedman test provides  $p_{value} = 0.00001$ , leading to reject  $H_0$ . The *pos hoc* obtained by Wilcoxon test is described in Table 11. According to the results, the classifier REC is significantly different when compared to the other



ones. In addition, the KNN classifier is significantly different from the LMC, LPC, HKNN and REC. Lastly, the FKNN is significantly different than LMC, HKNN, REC and ALH. Therefore, the statistical tests indicate that the best classifier is the REC and the worst classifiers are KNN and EKNN. Also in this case, this finding is consistent with the results obtained by A-TOPSIS.

## 5 Concluding Remarks

In this work, we present a thorough investigation about our previous work, the A-TOPSIS framework. We carried out two cases studies case in which we detailed the applicability of our approach and we compare it with the Hellinger-TOPSIS. Throughout the experiments, we described the benefits of using the A-TOPSIS rather the Hellinger-TOPSIS. In order to verify the suitability of the A-TOPSIS rank, we performed the nonparametric tests of Friedman and Wilcoxon. The obtained results showed the effectiveness of the approach and indicate that the A-TOPSIS can support the statistical tests with a complete rank of all algorithms analyzed.

Despite we use classification problems in both studies case, the presented approach is general and can be applied to compare the performance of any stochastic algorithms in machine learning. In terms of computational burden, the A-TOPSIS consists of a very simple computation procedure. It is worth to note that the TOPSIS is a well-established and reliable methodology, which guarantee the A-TOPSIS effectiveness. Finally, in order to encourage researchers and practitioners in the different areas of knowledge, especially in machine learning, to use the A-TOPSIS, we provided a web framework to rank algorithms in an easy way.

**Acknowledgements** A.G.C. Pacheco would like to thank the financial support of the Brazilian agency CAPES and R.A. Krohling thanks the financial support of the Brazilian agency CNPq under Grant No. 309161/2015-0.

## References

1. Behzadian M, Otaghsara SK, Yazdani M, Ignatius J (2012) A state-of the-art survey of TOPSIS applications. *Expert Syst Appl* 39:13051–13069
2. Brazdil PB, Soares C (2000) A comparison of ranking methods for classification algorithm selection. In: *European conference on machine learning*. Springer, Berlin, Heidelberg, pp 63–75
3. Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
4. Derrac J, Garcia S, Molina D, Herrera F (2011) A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm Evol Comput* 1:3–18
5. García S, Molina D, Lozano M, Herrera F (2009) A study on the use of nonparametric tests for analyzing the evolutionary algorithms' behaviour: a case study on the CEC'2005 special session on real parameter optimization. *J Heuristics* 15:617–644
6. Hwang CL, Yoon KP (1981) *Multiple attribute decision making methods and applications*. Springer, Berlin
7. Krohling RA, Campanharo VC (2011) Fuzzy TOPSIS for group decision making: a case study for accidents with oil spill in the sea. *Expert Syst Appl* 38:4190–4197
8. Krohling RA, Pacheco AGC (2015) A-TOPSIS—an approach based on TOPSIS for ranking evolutionary algorithms. In: *International conference on information technology and quantitative management (ITQM 2015)*, Rio de Janeiro, Brazil, *Procedia Computer Science*, vol 55, pp 308–317

9. Krohling RA, Lourenzutti R, Campos M (2015) Ranking and comparing evolutionary algorithms with Hellinger-TOPSIS. *Appl Soft Comput* 36:217–226
10. Kotthoff L (2014) Ranking algorithms by performance. In: *International conference on learning and intelligent optimization*. Springer, Cham, pp 16–20
11. Lourenzutti R, Krohling RA (2014) The Hellinger distance in multicriteria decision making: an illustration to the TOPSIS and TODIM methods. *Expert Syst Appl* 41:4414–4421
12. Pacheco AGC (2016) Aggregation of neural classifiers using Choquet integral with respect to a fuzzy measure (in Portuguese). Master's thesis. Federal University of Espírito Santo. <http://repositorio.ufes.br/handle/10/6811>
13. Peng Y, Kou G, Wang G, Shi Y (2011) FAMCDM: a fusion approach of MCDM methods to rank multiclass classification algorithms. *Omega* 39(6):677–689
14. Rice JR (1976) The algorithm selection problem. *Adv Comput* 15:65–118
15. Wen G, Chen X, Jiang L, Li H (2013) Performing classification using all kinds of distances as evidences. In: *12th IEEE international conference in cognitive informatics and cognitive computing*, pp 168–174