Comparing partitions through the Matching Error

Mathias Bourel, Badih Ghattas, Meliza González

Comparing partitions through the Matching Error

Mathias Bourel[*], Badih Ghattas[†], Meliza González [‡]

July 30, 2019

Abstract

With the aim to propose a non parametric hypothesis test, this paper carries out a study on the Matching Error (ME), a comparison index of two partitions obtained from the same data set, using for example two clustering methods. This index is related to the misclassification error in supervised learning. Some properties of the ME and, especially, its distribution function for the case of two independent partitions are analyzed. Extensive simulations show the efficiency of the ME and we propose a hypothesis test based on it.

1 Introduction

Most clustering approaches result in a partition of the data set and often a partition of the space where the data lie. Several indices may be used to compare partitions coming from a same data set, among which the Rand index (Rand (1971)), the Adjusted Rand Index (Hubert and Arabie (1985)), the Jaccard Index (Hultsch (2004)), etc. They can be also used to assess the performance of a clustering approach over a supervised dataset. Most of these existing indices lack real mathematical analysis, and almost no information exists about their distribution.

We consider here the Matching Error (ME) introduced by Meila (Meilă and Heckerman (2001), Meilă (2005)) and inspired by the classification error rate used in supervised learning. This index has been used in few works (for example and recently in Fraiman et al. (2013)) because its computation for large number of clusters is quite complex. We wish to derive in this paper a hypothesis test to compare two partitions, based on the ME statistic. For that, we focus on the theoretical properties of the ME, in particular to derive its distribution and show its efficiency in various experimental designs.

This paper is organized as follows. In Section 2, we present a state of the art of some classical indices for comparison of two partitions and their properties. Section 3 is devoted to the study of the ME where we establish some theoretical results, in particular its distribution in case of

[*]M. Bourel, mbourel@fing.edu.uy, IMERL, Facultad de Ingeniería, Universidad de la República, Julio Herrera y Reissig 565, 11300 Montevideo, Uruguay. Corresponding author

[†]B. Ghattas, badih.ghattas@univ-amu.fr, Institut de Mathématiques de Marseille, Université Aix-Marseille, France

[‡]M. González, meliza.gonzalez@gmai.com, Universidad de la República, Montevideo, Uruguay.

independence of partitions and balanced clusters. Several properties are proved for the general case. In Section 4, we show many simulations varying the experimental designs (sample size, number of groups, dependence of the partitions) to compare the ME with other indices. This section ends with our proposal of an hypothesis test for the independence of two partitions, designed from the theoretical results on ME, and its performance.

2 Related works

We denote by $\mathcal{L} = \{x_1, \ldots, x_n\} \subseteq \mathbb{R}^p$ a sample of n independent realizations of a multivariate random variable $X = (X_1, \ldots, X_p)$. Clustering seeks to form disjoint subgroups of observations such that individuals within the same cluster are similar to each other and relatively different from those of the other clusters. Let $\mathscr{C}$ be a partition of $\mathcal{L}$ obtained by a cluster analysis, that is $\mathscr{C}$ is a collection of disjoint subsets $\{C_1, \ldots, C_J\}$ such that their union is $\mathcal{L}$. The set of all possible partitions of $\mathcal{L}$ is denoted $\mathcal{P}(\mathcal{L})$. Let $\mathscr{C}' = \{C'_1, \ldots, C'_L\} \in \mathcal{P}(\mathcal{L})$ be a second partition of $\mathcal{L}$. The number of clusters of partitions $\mathscr{C}$ and $\mathscr{C}'$ (J and L respectively) may be different.

Following work of Wagner and Wagner (2007), there exist three kind of similarity (or dissimilarity) measures between two partitions; we give a brief review of such measures.

- Measures based on counting pairs

  A natural way to compare partitions is by counting pairs of observations belonging to a same cluster in both partitions. The set of all (unordered) pairs of $\mathcal{L}$ is the disjoint union of the following sets:

  - A={pairs of observations that are in the same cluster in $\mathscr{C}$ and $\mathscr{C}'$}
  - B={pairs of observations that are in different clusters in $\mathscr{C}$ and $\mathscr{C}'$}
  - C={pairs of observations that are in the same cluster in $\mathscr{C}$ but in different clusters in $\mathscr{C}'$}
  - D={pairs of observations that are in the same cluster in $\mathscr{C}'$ but in different clusters in $\mathscr{C}$}

  Sets A, B, C and D are disjoint and if $a = |A|, b = |B|, c = |C|$ and $d = |D|$, where $|\cdot|$ stands for the cardinal) we have $a + b + c + d = \frac{n(n-1)}{2}$.

  A very common index based on counting pairs is the *Rand index* (Rand (1971)) defined by:

$$R(\mathscr{C}, \mathscr{C}') = \frac{a+b}{a+b+c+d} = \frac{2(a+b)}{n(n-1)}$$

  It counts the proportion of pairs classified in a same way by the two clusterings. It is equal to zero when there exist no pairs of observations classified in the same way by both clustering, and it is equal to one when the two partitions are identical. Because the expected value of the Rand index of two random partitions is not constant, Hubert and Arabie (1985) proposed an adjustment based on the hypothesis that the clusterings are generated randomly subject

to a fixed number of groups and fixed cluster size. The *Adjusted Rand* index is a normalized version of the Rand index and is defined as:

$$R_{adj}\left(\mathscr{C},\mathscr{C}'\right)=\frac{R\left(\mathscr{C},\mathscr{C}'\right)-\mathbb{E}(R\left(\mathscr{C},\mathscr{C}'\right))}{\max(R\left(\mathscr{C},\mathscr{C}'\right))-\mathbb{E}(R\left(\mathscr{C},\mathscr{C}'\right))}$$

which is equivalent to:

$$R_{adj}\left(\mathscr{C},\mathscr{C}'\right)=\frac{a-((a+d)(a+c)/(a+b+c+d))}{\frac{(a+d)+(a+c)}{2}-\frac{(a+d)(a+c)}{a+b+c+d}}$$

Another indices are *Fowlkes-Mallows index* (Fowlkes and Mallows (1983)), *Mirkin metrix*(Mirkin (1998)), *Partition Difference* (Li et al. (2004)) and *Jaccard index* (Hultsch (2004)) . The latter measures the similarity between two partitions. It is very similar to the Rand index, but it dismisses the pairs of elements that are in different clusters in the compared partitions. It is defined as

$$J(\mathscr{C},\mathscr{C}')=\frac{a}{a+c+d}$$

However, many of these measures have undesirable properties such as sensitivity to the number of clusters, the number of observations and the relative size of clusters.

- Measure based on set overlaps

  Measurements based on set overlaps are in general computed from the confusion matrix between the two partitions $\mathscr{C}$ and $\mathscr{C}'$. The matrix $N=(n_{ij})\in J\times L$ is such that $n_{ij}=\left|C_i\cap C'_j\right|$, $1\leq i\leq J$, $1\leq j\leq L$. We will suppose that $J\leq L$.

  Meilă and Heckerman (2001) introduced an index called *the classification error* inspired from the misclassification error used in supervised learning. Consider that one of the two compared clusterings ($\mathscr{C}$ for instance) corresponds to the true labels of each observation and the other clustering ($\mathscr{C}'$) to the predicted ones. The supervised classification error may be computed for all the possible permutations of the predicted labels (in $\mathscr{C}'$), and the maximum error over all the permutations may be taken. Thus the classification error for comparing both partitions may be written as

$$CE(\mathscr{C},\mathscr{C}')=1-\frac{1}{n}\max_{\sigma}\sum_{i=1}^{J}n_{i\sigma(i)} \tag{1}$$

  where σ is an injective mapping of $\{1,\ldots,J\}$ into $\{1,\ldots,L\}$ (Meilă (2005)). The ME index may be complex to compute when the number of clusters is large. A polynomial time algorithm has been proposed by Fraiman et al. (2013) to compute it efficiently. We will study the distributional properties of this index in the next section.

- Measures based on mutual information

  The entropy of a partition $\mathscr{C}$ is defined by $H(\mathscr{C}) = -\sum_{i=1}^{J} p(i)\log_2 p(i)$ where $p(i) = |C_i|/n$ is the estimate of the probability that an element is in cluster $C_i \in \mathscr{C}$. The *mutual information* can be used to measure the independence of two partitions $\mathscr{C}$ and $\mathscr{C}'$. It is given by: $I(\mathscr{C},\mathscr{C}') = \sum_{i=1}^{J}\sum_{j=1}^{L} p(i,j)\log_2 \frac{p(i,j)}{p(i)p(j)}$, where $p(i,j)$ is the estimate of the probability that an element belongs to cluster C_i of $\mathscr{C}$ and C'_j of $\mathscr{C}'$. Mutual information is a metric over the space of all clusterings, but its value is not bounded which makes it difficult to interpret. As $I(\mathscr{C},\mathscr{C}') \leq \min(H(\mathscr{C}), H(\mathscr{C}'))$, other bounded indices have been proposed such as *Normalized Mutual Information* (Strehl and Ghosh (2002), Fred and Jain (2003)) where $I(\mathscr{C},\mathscr{C}')$ is divided either by the arithmetic or the geometric mean of the clustering entropies. Meila (Meilă (2003)) has also proposed an index based on Mutual information called *Variation of Information*.

In de Souto et al. (2012) and Rezaei and Fränti (2016) several indices are compared on artificially simulated partitions with various configurations; partitions are either balanced or unbalanced, dependent or independent, varying number of clusters. They show that the indices based on set overlaps have better performance than those based on counting pairs and mutual information. Besides, most indices are not relevant when the clusters in the partitions are imbalanced.

Milligan and Cooper (1986) study the behavior of the Rand, Adjusted Rand, Jaccard and Fowlkes Mallows indices. They compare the partitions produced by hierarchical algorithms with the true partitions, varying the number of groups with a sample of 50 observations and conclude that the adjusted Rand index seems to be more appropriate for clustering validation in this context. Similar simulations and results are given in Brun et al. (2007) and Wu et al. (2009) using k-means.

The works cited above give some experimental conclusions with no theoretical framework. In Saporta and Youness (2002), Youness and Saporta (2004) and Youness and Saporta (2010), the authors propose methods to study the empirical distribution of partition comparison indices, in particular the Rand and the Jaccard indices, among others. By a latent class mode to generate the data, they estimate their empirical distribution under the hypothesis that the two partitions come from the same underlying mixture model. Distributions of these indices depend on the number of clusters, their proportions and their separation so it is impossible to derive a general result.

Indices for comparing partitions should have some desirable properties, like being bounded, interpretable, independent of the number of clusters and sample size, and complying with properties of a metric. In Meilă (2005), the author makes an axiomatic characterization of Variation of information, Mirkin, Rand and Van Dongen indices and some properties of the ME are discussed, in particular that it is a metric in some subspaces of the clusterings sets. In Meilă (2007) numerous theoretical properties on the variation of information index are proved. Also, it is shown that the normalized Mirkin metric and the Adjusted Rand index satisfy the properties of a metric.

With respect to the theoretical distribution of the indices, Idrissi (2000) established the following statement: for two independent partitions, with J balanced clusters, the asymptotic dis-

tribution of the Rand Index is normal with expectation $\mathbb{E}\left(R\left(\mathscr{C},\mathscr{C}'\right)\right)=1-\frac{2}{J}+\frac{2}{J^2}$ and variance $\mathbb{V}\left(R\left(\mathscr{C},\mathscr{C}'\right)\right)=\frac{1}{n^2}\left(1-\frac{1}{n}\right)\left(1-\frac{2}{J}+\frac{2}{J^2}\right)\left(\frac{2}{J}-\frac{2}{J^2}\right)$. But this result is not true for small J, in particular when $J=2$, and is only approximately valid only for large samples. This is, to our best knowledge, the only theoretical result about the distribution of the Rand index and under some conditions. Our work intends to analyze the distribution of the ME index.

3 The ME properties

We now study the distributional properties of the classification error introduced in (1) that we formulate in an equivalent way. More precisely, let us consider a data set $\mathcal{L}=\{x_1,\ldots,x_n\}\subseteq\mathbb{R}^p$ and $\mathscr{C}$ and $\mathscr{C}'\in\mathcal{P}(\mathcal{L})$ two partitions of $\mathcal{L}$. The labels of each observation in the first partition are denoted $\{y_1,\ldots,y_n\}$ and those of the second partition $\{\hat{y}_1,\ldots,\hat{y}_n\}$, so $y_i\in\{1,\ldots,J\}$ and $\hat{y}_i\in\{1,\ldots,L\}\,\forall i=1,\ldots,n$.

If S_J is the set of permutations of $\{1,\ldots,J\}$ and $\mathcal{A}$ is the set of arrangements of J elements taken from L, we define the *Matching Error* (ME) as:

$$\tau=ME(\mathscr{C},\mathscr{C}')=\begin{cases}\min\limits_{\sigma\in S_J}\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{1}_{\{y_i\neq\sigma(\hat{y}_i)\}} & \text{if } J\leq L\\ \min\limits_{\sigma\in\mathcal{A}}\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{1}_{\{\sigma(y_i)\neq\hat{y}_i\}} & \text{otherwise}\end{cases}\tag{2}$$

For simplicity, we assume that the two compared partitions have the same number of clusters, that is $J=L$. Observe that $ME(\mathscr{C},\mathscr{C}')$ is another formulation of $CE(\mathscr{C},\mathscr{C}')$ used in the works of Meilă and we are going to analyze its distribution.

3.1 Distribution function of τ_σ

Let denote $\tau_\sigma=\frac{1}{n}\sum\limits_{i=1}^{n}\mathbf{1}_{\{y_i\neq\sigma(\hat{y}_i)\}}$. To derive the distribution of the ME we need to know the distribution of the different classification errors τ_σ where $\sigma\in S_J$. For y_i and $\hat{y}_i$ two independent realizations of a discrete random variable Y taking values in $\{1,\ldots,J\}$, let:

- $p_j=\mathbb{P}[y_i=j]$, $\hat{p_j}=\mathbb{P}[\hat{y}_i=j]$, and
- $\theta=\mathbb{P}[y_i\neq\sigma(\hat{y}_i)]=1-\mathbb{P}[y_i=\sigma(\hat{y}_i)]$
  $=1-\sum\limits_{j=1}^{J}\mathbb{P}[y_i=j,\sigma(\hat{y}_i)=j]=1-\sum\limits_{j=1}^{J}p_j\hat{p}_j$

Assuming that Y is uniform on $\{1,\ldots,J\}$, then $p_j=\hat{p}_j=1/J$ and $\theta=1-\frac{1}{J}$. Therefore the random variable $n\tau_\sigma$ is binomial with parameters n and θ, and:

$$\mathbb{E}(\tau_\sigma)=1-\frac{1}{J};\quad\text{and}\quad\mathbb{V}(\tau_\sigma)=\frac{1}{n}\frac{1}{J}\left(1-\frac{1}{J}\right)$$

For large values of n: $n\tau_\sigma\sim\mathcal{N}\left(n\theta,n\theta(1-\theta)\right)$.

Proposition 1. *1. As* $\sum_{j=1}^{J!} n\tau_{\sigma_j} = n(J-1)!(J-1)$*, the random variables* $n\tau_{\sigma_1}, n\tau_{\sigma_2}, \ldots, n\tau_{\sigma_{J!}}$ *are not independent.*

2. The ME is bounded:

$$0 \leq \tau \leq \frac{J-1}{J}$$

Let σ_l and σ_k be two permutations of S_J. We say that σ_l and σ_k share a point $j \in \{1, \ldots, J\}$ if $\sigma_l(j) = \sigma_k(j)$. Note that two permutations of S_J can share at most $J-2$ points.

Proposition 2. *If* σ_j *and* σ_l *share* s *points, then*

$$\mathbb{COR}(n\tau_{\sigma_l}, n\tau_{\sigma_k}) = \frac{s-1}{J-1}$$

where $s = 0, \ldots, J-2$

In Figure 1 we show the distribution of τ for different values of J and for $n = 1000$.

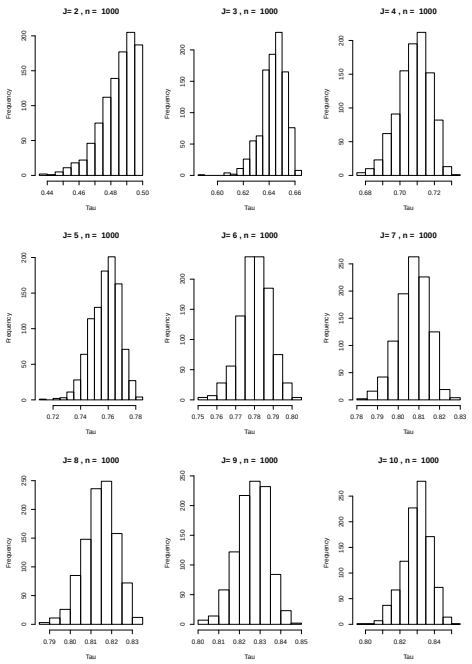

Figure 1: Distribution of τ for $n = 1000$ and different values of J for two independent and balanced partitions.

3.2 The case $J = 2$

Proposition 3. *The distribution function* $F_{n\tau}(z)$ *of* $n\tau = \min\{n\tau_{\sigma_1}, n\tau_{\sigma_2}\}$ *is:*

Table 1: Correlation between $ME(\mathscr{C},\mathscr{C}')$ and $R(\mathscr{C},\mathscr{C}')$ over $N = 1000$ repetitions of the indices, where $\mathscr{C}$ and $\mathscr{C}'$ are partitions with $J = 2,\dots,10$ groups and $n = 1000$ observations in the four scenarios explained above.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	mean
Dependence	Unbalanced groups	-1	-0,93	-0,89	-0,83	-0,81	-0,79	-0,73	-0,74	-0,70	-0,83
	Balanced groups	-1	-1	-1	-1	-0,99	-0,99	-0,99	-0,99	-0,98	-0,99
Independence	Unbalanced groups	-0,94	-0,28	-0,43	-0,21	-0,21	-0,12	-0,14	-0,13	-0,05	-0,28
	Balanced groups	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

$$\mathbb{P}(n\tau \le z) = \begin{cases} \sum_{i=0}^{z} 2\binom{n}{i}\left(\frac{1}{2}\right)^n = 2\mathcal{B}_{n,1/2}(z) & \text{if } z+1 \le \frac{n}{2} \\ 1 & \text{if } z > \frac{n}{2} \end{cases}$$

where $B_{n,1/2}$ is the binomial distribution function with parameters (n, $\frac{1}{2}$).

Proposition 4.

$$\mathbb{E}(n\tau) = \begin{cases} \frac{n}{2} - \left(\frac{1}{2}\right)^n \binom{n}{n/2}\frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n}{2} - \left(\frac{1}{2}\right)^n \binom{n-1}{\frac{n-1}{2}} n & \text{if } n \text{ is odd} \end{cases}$$

Hence the distribution of ME is explicit for $J = 2$.

4 Simulations and hypothesis test

We first show how the ME behaves in different experimental situations and derive a hypothesis test for independence of two partitions.

4.1 Empirical performance of the ME

In this section we study the empirical distribution of ME between two partitions $\mathscr{C}$ and $\mathscr{C}'$. The results are considered and compared in various experimental conditions: different number of groups ($J \in \{2,3,4,5,6,7,8,9,10\}$), different number of observations ($n \in \{50,100,200,300,400,500,1000\}$), independent partitions, different degrees of dependence between partitions, balanced and imbalanced clusters.
For each configuration of these parameters, we directly generate the partition vectors, that is, each observation's label for both partitions. These could be the result of two clustering analysis. For each value of J and n, we generate $N = 1000$ independent partition pairs and the index $\tau = ME(\mathscr{C},\mathscr{C}')$ is computed. This produces 1000 values of the index for each configuration.

To simulate dependent partitions, we start with two equal vectors Y and $\hat{Y}$ and modify at random a proportion $\gamma \in \{0.1, 0.4, 0.6, 0.9\}$ of labels of $\hat{Y}$.
We check first the correlations between the ME, and both the Rand and Jaccard indices. Tables 1 and 2 give the obtained results for the four scenarios, for $\gamma = 0.4$.

In the case of balanced clusters, the correlation is high (it is negative as ME is a dissimilarity measure between partitions, whereas Rand and Jaccard indices are similarity measures). For dependent partitions the correlation is almost perfect (close to minus one), while with independent

Table 2: Correlation between $ME(\mathscr{C},\mathscr{C}')$ and Jaccard$(\mathscr{C},\mathscr{C}')$ over $N = 1000$ repetitions of the indices, where $\mathscr{C}$ and $\mathscr{C}'$ are partitions with $J = 2,\ldots,10$ groups and $n = 1000$ observations in the four scenarios explained above.

		J=2	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10	mean
Dependence	Unbalanced groups	-0,99	-0,95	-0,91	-0,87	-0,85	-0,82	-0,78	-0,77	-0,75	-0,86
	Balanced groups	-1	-1	-1	-1	-1	-1	-1	-1	-1	-0,99
Independence	Unbalanced groups	-0,58	-0,45	-0,6	-0,52	-0,56	-0,59	-0,57	-0,63	-0,54	-0,56
	Balanced groups	-0,93	-0,86	-0,83	-0,81	-0,8	-0,79	-0,78	-0,78	-0,77	-0,82

partitions it is still high but smaller (between -0.77 and -0.93). Correlation between the indices decreases when the number of clusters J increases.
For unbalanced partitions results are quite different. For dependent partitions, the correlations -although decreasing with respect to the cluster size- have a coefficient between -0.7 and -0.99. However, when the partitions are independent, the correlation is smaller and, except for the case $J = 2$, it is much weaker with the Rand index than with the Jaccard index.

4.2 Hypothesis Test

Our main purpose when analysing the distribution of the ME index is to design a hypothesis test to decide whether two partitions are statistically independent. The properties proved above were derived under some assumptions and may be used to compare partitions at least for $J = 2$ groups equally distributed. We present this test and analyse its performance on simulated data.

Given two partitions $\mathscr{C}$ and $\mathscr{C}'$, the test proposal is:

(H_0): Partitions $\mathscr{C}$ and $\mathscr{C}'$ are independent.

(H_1): Partitions $\mathscr{C}$ and $\mathscr{C}'$ are not independent.

The test statistic is τ and, as we know its distribution for $J = 2$ under H_0 it is straightforward to derive decision.

For $J > 2$, we propose the following reasoning. For two partitions $\mathscr{C}$ and $\mathscr{C}'$ we compute $\tau_0 = ME(\mathscr{C},\mathscr{C}')$ and then take B "perturbations" of $\mathscr{C}'$ denoted $\mathscr{C}'_1,\mathscr{C}'_2,\ldots,\mathscr{C}'_B$, changing at random a proportion π of its labels. To estimate the p-value $\mathbb{P}(\tau \leq \tau_0)$ of the independence hypothesis (null hypothesis) we take the proportion of values of $\tau_b = ME(\mathscr{C},\mathscr{C}'_b)$ which are less than τ_0. We give two examples of the use of this test, for $J = 15$. First we consider two independent balanced partitions of $n = 1000$ observations with $J = 15$ groups and fix $B = 1000$. In this case $\tau_0 = 0.881$ and the estimated $p-$value is 0.464, which is coherent with not rejecting the hypothesis of independence of the two partitions. In Figure 2 we show the histogram of the B different values of τ. In the second case and for the same values of n and J, we consider two dependent partitions where the second is constructed from the first by changing 20% of the labels randomly. In this case $\tau_0 = 0.186$ and the estimated p-value equals 0 so we reject the hypothesis of dependence. In Figure 3 we show the histogram of the B different values of τ.

To evaluate the performance of the test, we calculate the error averaging over $N = 1000$ simulations. We take $B = 1000$ and proportion π equals 0.2. At level $\alpha = 0.1$, we estimate the number of times we make type I and type II error using the empirical values of the p-value. As it was expected, taking proportion of values of p-value less than 0.1, the estimation of doing a type I error is small, less than 8%, when we compare two independent partitions. We give in Table 3

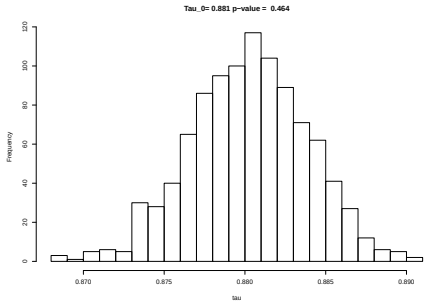

Figure 2: Histogram of the $B = 1000$ values of τ obtained by changing 20% of the labels randomly with two independent partitions.

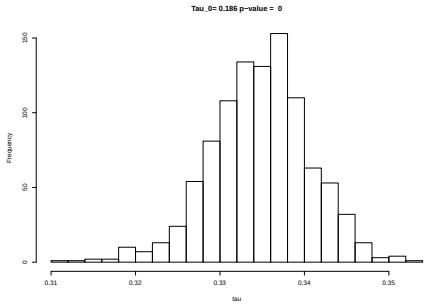

Figure 3: Histogram of the $B = 1000$ values of τ with two dependent partitions.

the individual values for some n and varying $J \in \{3, \ldots, 10\}$.

Table 3: At level 0.1, proportion of times a type I error is made obtained by averaging $N = 1000$ comparisons of independent partitions with balanced classes.

	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	1.1	1.6	3.3	3.1	4.3	6.1	7.9	7.1
n=100	0.3	1.2	1.6	3.1	3.6	3.1	4.0	4.5
n=200	0.2	0.8	1.2	1.5	1.7	2.5	2.8	3.0
n=300	0.4	0.7	1.1	1.3	1.6	1.1	2.4	2.3
n=400	0.2	0.4	0.6	1.3	1.8	1.3	1.4	1.7
n=500	0.0	0.2	1.1	0.9	1.2	1.7	1.4	2.8
n=1000	0.0	0.1	0.8	0.9	0.8	1.4	1.4	2.2

On the other hand, with dependent partitions, to estimate the proportion of type II errors over the N simulations, we take the proportion of p-values larger than 0.1. As described in the previous section, partitions with balanced clusters and different degrees of dependence are simulated where dependency strength γ varies in $\{0.1, 0.4, 0.6, 0.9\}$. This estimation equals zero when $\gamma = 0.1$ and $\gamma = 0.4$ accordingly with a high dependence of the partitions. For $\gamma = 0.6$ it is null except for $n = 50$: for $J = 3$ it equals 17.8%, and it is less than 3% only for $4 \leq J \leq 10$. When $\gamma = 0.9$, which is a scenario very close to independence, it has high values for $n = 50, 100$ an 200 but decreases to 0 when n and J grow as we can see in Table 4.

Table 4: At level 0.1, proportion of times a type II error is made obtained averaging $N = 1000$ comparisons of dependent partitions with $\gamma = 0.9$, with balanced classes.

	J=3	J=4	J=5	J=6	J=7	J=8	J=9	J=10
n=50	98.3	97.4	96.0	95.3	94.4	93.3	92.2	91.9
n=100	97.7	96.4	95.1	93.6	91.1	91.3	90.4	91.8
n=200	94.3	89.0	87.7	83.0	83.1	82.5	79.2	82.8
n=300	89.5	78.9	72.7	65.9	67.5	63.9	63.2	62.7
n=400	81.8	63.0	53.4	46.1	43.1	40.8	40.5	38.1
n=500	73.1	46.5	35.0	27.9	21.5	23.4	20.3	17.3
n=1000	22.2	3.7	1.6	0.2	0.2	0.0	0.0	0.0

The experimental evaluation shows a good performance of the partition independence test with balanced clusters. The proportion of times for which a type I error occurs is in the expected environment for a level test with $\alpha = 0.1$. In the evaluation of type II error, the test proved to be powerful when the dependence between the partitions is relatively high and that it loses precision when the compared partitions have conditions that approximate the null hypothesis of independence.

5 Conclusions

We have suggested an hypothesis test for comparing two partitions, useful for comparing the results of two clustering approaches over a same dataset. Our test is based on the mismatch error inspired from the misclassification error in supervised learning. We have analyzed the properties and the distribution of this index in several conditions and compared it to other common indices. A closed form of the statistic distribution under the null hypothesis was given for two clusters under mild conditions. For more than two clusters, the simulations show that the test is quite robust and reliable in various experimental conditions, but the statistic distribution under the null hypothesis is still unavailable.

Appendix

Proof of Proposition 1

1. $n\sum_{j=1}^{J!}\tau_{\sigma_j} = n\frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{\{y_i\neq\sigma_1(\hat{y}_i)\}}+\cdots+\mathbf{1}_{\{y_i\neq\sigma_{J!}(\hat{y}_i)\}}\right) = \sum_{i=1}^{n}|\{\sigma\in S_J:\sigma(\hat{y}_i)\neq y_i\}|$
   For fixed y_i and $\hat{y}_i$ we have: $|\{\sigma\in S_J:\sigma(a)\neq b\}| =$
   $|S_J| - |\{\sigma\in S_J:\sigma(a)=b\}| = J! - (J-1)! = (J-1)!(J-1)$
   it follows that $n\sum_{j=1}^{J!}\tau_{\sigma_j} = n(J-1)!(J-1)$.

2. It is straightforward, because if τ was greater than $\frac{J-1}{J}$, then $\sum_{j=1}^{J!}\tau_{\sigma_j} > J!\frac{J-1}{J} = (J-1)!(J-1)$,

which contradicts part 1.

Proof of Proposition 2

$$\begin{aligned}
&\mathbb{COV}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) = \mathbb{E}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) - \mathbb{E}(n\tau_{\sigma_k})\mathbb{E}(n\tau_{\sigma_l}) \\
=&\mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_k(\hat{y}_i)\}} \sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_l(\hat{y}_j)\}}\right) - \mathbb{E}\left(\sum_{i=1}^{n} \mathbf{1}_{\{y_i \neq \sigma_k(\hat{y}_i)\}}\right) \mathbb{E}\left(\sum_{j=1}^{n} \mathbf{1}_{\{y_j \neq \sigma_l(\hat{y}_j)\}}\right) \\
=&\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{P}\big(y_i \neq \sigma_k(\hat{y}_i), y_j \neq \sigma_l(\hat{y}_j)\big) - n^2\big(1 - 1/J\big)^2
\end{aligned}$$

Using De Morgan equalities and basic probability properties, it is easy to show that, by independence $\mathbb{P}\left(y_i = \sigma_k(\hat{y}_i), y_j = \sigma_l(\hat{y}_j)\right)$

$$= \begin{cases} 1/J^2 \text{ if } i \neq j \\ \sum_{j=1}^{J} \mathbb{P}\left(y_i = j\right) \underbrace{\mathbb{P}\left(\sigma_k(\hat{y}_i) = j, \sigma_l(\hat{y}_i) = j\right)}_{(*)} \text{ if } i = j \end{cases}$$

If σ_k and σ_l share s points, $(*) = \frac{s}{J^2}$ and therefore

$$\mathbb{COV}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) = \frac{n(s-1)}{J^2}$$

and

$$\mathbb{COR}(n\tau_{\sigma_k}, n\tau_{\sigma_l}) = \frac{\frac{n(s-1)}{J^2}}{n\frac{1}{J}\left(1 - \frac{1}{J}\right)} = \frac{s-1}{J-1}$$

Proof of Proposition 3 As $n\tau_{\sigma_1} + n\tau_{\sigma_2} = n$, if $n\tau_{\sigma_1} = x$ and $n\tau_{\sigma_2} = y$, it is easy to establish that:

$$\mathbb{P}(n\tau_{\sigma_1} = x, n\tau_{\sigma_2} = y) = \begin{cases} 0 & \text{if } x + y \neq n \\ \binom{n}{x}\left(\frac{1}{2}\right)^n & \text{if } x + y = n \end{cases}$$

and the joint probability table for $n\tau_{\sigma_1}$ and $n\tau_{\sigma_2}$ has the following shape:

$n\tau_{\sigma_1}/n\tau_{\sigma_2}$	0	1	2	$\cdots$	$n-2$	$n-1$	n
0	0	0	0	0	0	0	$\binom{n}{0}\left(\frac{1}{2}\right)^n$
1	0	0	0	0	0	$\binom{n}{1}\left(\frac{1}{2}\right)^n$	0
2	0	0	0	0	$\binom{n}{2}\left(\frac{1}{2}\right)^n$	0	0
$\vdots$	0	0	0	$\cdots$	0	0	0
$n-2$	0	0	$\binom{n}{n-2}\left(\frac{1}{2}\right)^n$	0	0	0	0
$n-1$	0	$\binom{n}{n-1}\left(\frac{1}{2}\right)^n$	0	0	0	0	0
n	$\binom{n}{n}\left(\frac{1}{2}\right)^n$	0	0	0	0	0	0

.

Using this table, we have;

$$\begin{aligned}
&\mathbb{P}(n\tau = z)\\
&\quad = \mathbb{P}(n\tau_{\sigma_1} = z, n\tau_{\sigma_2} = n - z) + \mathbb{P}(n\tau_{\sigma_2} = z, n\tau_{\sigma_1} = n - z)\\
&\quad = 2\mathbb{P}(n\tau_{\sigma_1} = z,\, n\tau_{\sigma_2} = n - z)\\
&\quad = 2\binom{n}{z}\left(\frac{1}{2}\right)^n, \text{ if } z \neq \frac{n}{2}.
\end{aligned}$$

If n is even:

$$\mathbb{P}\left(n\tau = \frac{n}{2}\right) = \mathbb{P}\left(n\tau_{\sigma_1} = \frac{n}{2}, n\tau_{\sigma_2} = \frac{n}{2}\right) = \binom{n}{n/2}\left(\frac{1}{2}\right)^n$$

So we can conclude that

$$\mathbb{P}(n\tau = z) = \begin{cases} \binom{n}{n/2}\left(\frac{1}{2}\right)^n & \text{if } z = \frac{n}{2} \text{ and } n \text{ is even} \\ 2\binom{n}{z}\left(\frac{1}{2}\right)^n & \text{otherwise} \end{cases}$$

Then, the distribution function of $n\tau$ is:

$$\begin{aligned}
F_{n\tau}(z) =& \mathbb{P}(n\tau \leq z) = \mathbb{P}(\min\{n\tau_{\sigma_1}, n\tau_{\sigma_1}\} \leq z)\\
=& 1 - \mathbb{P}(\min\{n\tau_{\sigma_1}, n\tau_{\sigma_1}\} \geq z + 1)\\
=& 1 - \mathbb{P}(n\tau_{\sigma_1} \geq z + 1, n\tau_{\sigma_2} \geq z + 1)\\
=& \mathbb{P}(n\tau \leq z) = \begin{cases} \sum_{i=0}^{z} 2\binom{n}{i}\left(\frac{1}{2}\right)^n = 2B_{n,1/2}(z) & \text{if } z + 1 \leq \frac{n}{2} \\ 1 & \text{if} z > \frac{n}{2} \end{cases}
\end{aligned}$$

Proof of Proposition 4

- Based on the table of joint probability, if n is even
  $\mathbb{E}(n\tau) = \sum_{k=0}^{n/2-1} 2\left(\frac{1}{2}\right)^n k\binom{n}{k} + \frac{n}{2}\left(\frac{1}{2}\right)^n \binom{n}{n/2}$.
  The first term is:
  $2\left(\frac{1}{2}\right)^n \sum_{k=1}^{n/2-1} k\binom{n}{k} = \left(\frac{1}{2}\right)^{n-1} \sum_{k=1}^{n/2-1} kn\frac{\binom{n-1}{k-1}}{k} = \left(\frac{1}{2}\right)^{n-1} n \sum_{k=1}^{n/2-1} \binom{n-1}{k-1}$
  $= \left(\frac{1}{2}\right)^{n-1} n \sum_{k=0}^{n/2-2} \binom{n-1}{k} = \left(\frac{1}{2}\right)^{n-1} n\left(\frac{2^{n-1}}{2} - \binom{n-1}{n/2-1}\right)$,
  as $\sum_{k=0}^{n-1}\binom{n-1}{k} = \sum_{k=0}^{n/2-1}\binom{n-1}{k} + \sum_{k=n/2}^{n-1}\binom{n-1}{k} = 2^{n-1}$
  and $\sum_{k=0}^{n/2-1}\binom{n-1}{k} = \sum_{k=n/2}^{n-1}\binom{n-1}{k} = \frac{2^{n-1}}{2}$.
  We have now,
  $\mathbb{E}(n\tau) = \left(\frac{1}{2}\right)^{n-1} n\left(\frac{2^{n-1}}{2} - \binom{n-1}{n/2-1}\right) + \frac{n}{2}\left(\frac{1}{2}\right)^n \binom{n}{n/2} =$

$\frac{n}{2} - \left(\frac{1}{2}\right)^{n-1} n\binom{n-1}{n/2-1} + \frac{n}{2}\left(\frac{1}{2}\right)^{n}\binom{n}{n/2} =$

$\frac{n}{2} - \left(\frac{1}{2}\right)^{n-1}\binom{n}{n/2}\frac{n}{2} + \left(\frac{1}{2}\right)^{n}\binom{n}{n/2}\frac{n}{2} = \frac{n}{2} - \left(\frac{1}{2}\right)^{n}\binom{n}{n/2}\frac{n}{2}.$

- Similarly if n is odd:

$\mathbb{E}(n\tau) = \sum_{k=0}^{\frac{n-1}{2}} 2\left(\frac{1}{2}\right)^{n} k\binom{n}{k} = \left(\frac{1}{2}\right)^{n-1}\sum_{k=0}^{\frac{n-1}{2}} k\binom{n}{k}$

$= \left(\frac{1}{2}\right)^{n-1}\sum_{k=0}^{\frac{n-1}{2}} n\binom{n-1}{k-1} = \left(\frac{1}{2}\right)^{n-1} n\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k}$

$= \left(\frac{1}{2}\right)^{n-1} n\left[2^{n-2} - \frac{1}{2}\binom{n-1}{\frac{n-1}{2}}\right]$, as

$\sum_{k=0}^{n-1}\binom{n-1}{k} = \sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} + \binom{n-1}{\frac{n-1}{2}} + \sum_{k=\frac{n-1}{2}+1}^{n-1}\binom{n-1}{k} =$

$2\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} + \binom{n-1}{\frac{n-1}{2}} = 2^{n-1}$ Then: $\sum_{k=0}^{\frac{n-1}{2}-1}\binom{n-1}{k} = 2^{n-2} - \frac{1}{2}\binom{n-1}{\frac{n-1}{2}}.$

Hence,

$$\Rightarrow \mathbb{E}(n\tau) = \begin{cases} \frac{n}{2} - (\frac{1}{2})^{n}\binom{n}{n/2}\frac{n}{2} & \text{if } n \text{ is even} \\ \frac{n}{2} - (\frac{1}{2})^{n}\binom{n-1}{\frac{n-1}{2}} n & \text{if } n \text{ is odd} \end{cases}$$

Acknowledgments

We thank the *LIA-IFUM*, the *ANII* -Uruguay and Campus France for their financial support.

References

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., Dougherty, E.R., 2007. Model-based evaluation of clustering validation measures. Pattern recognition 40, 807–824.

Fowlkes, E.B., Mallows, C.L., 1983. A method for comparing two hierarchical clusterings. Journal of the American statistical association 78, 553–569.

Fraiman, R., Ghattas, B., Svarc, M., 2013. Interpretable clustering using unsupervised binary trees. Adv. Data Analysis and Classification 7, 125–145.

Fred, A.L., Jain, A.K., 2003. Robust data clustering., in: Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. pp. 128–136.

Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of classification 2, 193–218.

Hultsch, L., 2004. Untersuchung zur besiedlung einer sprengfläche im pockautal durch die tiergruppen heteroptera (wanzen) und auchenorrhyncha(zikaden). Studienarbeit TU Bergakademie Freiberg, Studiengang Geoökologie .

Idrissi, A.N., 2000. Contribution à l'unification de critères d'association pour variables qualitatives. Ph.D. thesis. Université Paris 6, France.

Li, T., Ogihara, M., Ma, S., 2004. On combining multiple clusterings, in: Proceedings of the thirteenth ACM international conference on Information and knowledge management, ACM. pp. 294–303.

Meilă, M., 2003. Comparing clusterings by the variation of information, in: Schölkopf, B., Warmuth, M.K. (Eds.), Learning Theory and Kernel Machines, Springer Berlin Heidelberg. pp. 173–187.

Meilă, M., 2005. Comparing clusterings: an axiomatic view, in: Proceedings of the 22nd international conference on Machine learning, ACM. pp. 577–584.

Meilă, M., 2007. Comparing clusterings-an information based distance. Journal of multivariate analysis 98, 873–895.

Meilă, M., Heckerman, D., 2001. An experimental comparison of model-based clustering methods. Machine learning 42, 9–29.

Milligan, G.W., Cooper, M.C., 1986. A study of the comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research 21, 441–458.

Mirkin, B., 1998. Mathematical classification and clustering: From how to what and why, in: Classification, data analysis, and data highways. Springer, pp. 172–181.

Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66. URL: http://gen.lib.rus.ec/scimag/index.php?s=10.2307/2284239, doi:10.2307/2284239.

Rezaei, M., Fränti, P., 2016. Set matching measures for external cluster validity. IEEE Transactions on Knowledge and Data Engineering 28, 2173–2186.

Saporta, G., Youness, G., 2002. Comparing two partitions: Some proposals and experiments, in: Compstat, Springer. pp. 243–248.

de Souto, M.C., Coelho, A.L., Faceli, K., Sakata, T.C., Bonadia, V., Costa, I.G., 2012. A comparison of external clustering evaluation indices in the context of imbalanced data sets, in: Neural Networks (SBRN), 2012 Brazilian Symposium on, IEEE. pp. 49–54.

Strehl, A., Ghosh, J., 2002. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. Journal of machine learning research 3, 583–617.

Wagner, S., Wagner, D., 2007. Comparing clusterings: an overview. Universität Karlsruhe, Fakultät für Informatik Karlsruhe.

Wu, J., Chen, J., Xiong, H., Xie, M., 2009. External validation measures for k-means clustering: A data distribution perspective. Expert Systems with Applications 36, 6050–6061.

Youness, G., Saporta, G., 2004. Some measures of agreement between close partitions. Student 51, 1–12.

Youness, G., Saporta, G., 2010. Comparing partitions of two sets of units based on the same variables. Advances in data analysis and classification 4, 53–64.