Empirical Study of Partitions Similarity Measures

Abdelkrim Alfalah, Lahcen Ouarbya and John Howroyd

Abstract—This paper compares four existing distances and similarity measures between partitions. The partition measures considered in this paper are the *Rand Index* (**RI**), the *Adjusted Rand Index* (**ARI**), the *Variation of Information* (**VI**) and finally, the *Normalised Variation of Information* (**NVI**). This work investigates the ability of these partition measures to capture three predefined intuitions: the variation within randomly generated partitions, the sensitivity to small perturbations and finally the independence from the dataset scale . It has been shown that the *Adjusted Rand Index* (**ARI**) performed well overall, regarding these three intuitions.

Keywords—Clustering, comparing partitions, similarity measure, partition distance, partition metric, similarity between partitions, clustering comparison

I. Introduction

A large number of clustering algorithms has been studied [1]. Many questions have arised as a result of this extensive research: how to compare the similarities between the resulting partitions [2], [3], [4], [5], [6], [7], [8], [9], [10], how to validate the clustering results [11], how to combine different partitions to generate a better single partition [12], [13] and how to use a clustering similarity measure for feature selection in high dimensional datasets [14]. Having a distance or a similarity measure on the space of partitions is beneficial in answering such questions. In [15] Wagner et al. have proposed a sub-division of such similarity measures between partitions into three categories: in the first category, the measure is based on counting of pair of elements classified in the same way in both partitions, *Rand Index* (**RI**) similarity measure [16] is a prominent example from this category. In the second category, the measure is based on summation of set overlaps such as *Van Dongen-Measure* [15]. Finally, in the third category, the distance measure focuses on mutual information based on the concept of entropy in information theory, an example of this category is the *Variation of Information* (**VI**) distance measure [18].

The main contribution of this paper is to introduce three experimental studies investigating several distance and similarity measures in their ability to capture three intuitions : variation, sensitivity and scalability of the dataset.

The rest of this paper is structured as follows: Section II consists of an overview of the main existing partition distance and similarity measures. The framework used to compare these measures is introduced in Section III. In Section IV an empirical study as well as a result discussion has been conducted to compare the performance of four existing distance and similarity measures : *Rand Index* (**RI**), *Adjusted Rand Index* (**ARI**) and the *Normalised Variation of Information* (**NVI**) under the three experiments defined in Section III. Finally, a conclusion and a direction of future work is given in Section V.

Lahcen Ouarbya in the Department of Computer Science, Goldsmiths College, University of London,
New Cross, London, London SE14 6NW, UK
L.Ouarbya@gold.ac.uk

II. Overview of existing Partition Distance Measures

A. Definitions and Notations

Let X be a finite set containing N elements, $|X|$ = N. A partition C is a set $\{C_1, ..., C_k\}$ of non-empty disjoint subsets of X such that their union equals X. We assume $|C_i| > 0$ for all $i = 1, ..., k$. Let $C' = \{C'_1, ..., C'_l\}$ denote a second partition of X. The confusion matrix $M = (m_{ij})$ of the pair (C, C') is a $k \times l$ matrix whose ij-th entry equals the number of elements in the intersection of the clusters C_i and C'_j

$$m_{ij} = \left|C_i \cap C'_j\right|, 1 \leq i \leq k \; and \; 1 \leq j \leq l \tag{1}$$

The partition C' is a refinement of C (C is a coarsening of C'), if each class of C' is contained in a class of C. Hence, by definition, C' has to satisfy the following equation:

$$\forall C'_j \in C', \exists C_i \in C \; such \; that \; C'_j \subseteq C_i \tag{2}$$

The coarsest common refinement of the two partitions C, C' is represented by $C \times C'$ and it's defined as follows :

$$C \times C' = \{C_i \cap C'_j \mid C_i \in C \; , C'_j \in C' \; and \; C_i \cap C'_j \neq \emptyset\} \tag{3}$$

The set of all unordered pairs of elements of X is the disjoint union of the following four sets:
S_{11} is the set of pairs that are in the same class under C and C';
S_{00} is the set of pairs that are different clusters under C and C';
S_{10} is the set of pairs that are in the same class under C, but in different ones under C'; and
S_{01} is the set of pairs that are in different clusters under C, but in the same under C'

The cardinality of the set S_{ij} is represented by n_{ij} where

$$n_{ij} = |S_{ij}|, \; i, j \in \{0, 1\}$$

The remaining of this section consists of a detailed description of some of the existing similarity and distance measures. *Rand Index* (**RI**) [16], *Adjusted Rand Index* (**ARI**) [17], the *Variation of Information* (**VI**) [18], and the *Normalized Variation of Information* (**NVI**)[18].

B. Rand Index (RI)

In [16] Rand proposed, *Rand Index* (**RI**), an index of similarity between two partitions, C and C', based on pairs of elements of the set X, where X a finite set of all elements. The **RI** similarity measure calculates the fraction of the pairs classified simultaneously in the same class (n_{11}) and those classified in different class under both partitions (n_{00}) to the total number of pairs.

Considering the notations given in Section II-A, the *Rand Index* similarity measure is given as below :

$$\mathbf{RI}(C, C') = \frac{n_{00} + n_{11}}{n_{11} + n_{10} + n_{01} + n_{00}} \tag{4}$$

where $n_{i,j}$ is defined in SectionII-A

C. Adjusted Rand Index (ARI)

The *Rand Index* (**RI**) as a similarity measure has some drawbacks: in the case of two randomly generated partitions, its expected value is not constant. To overcome this limitation, Hurbert and Arabie [17] proposed a new similarity index measure, called the *Adjusted Rand Index* (**ARI**). The **ARI** similarity index can be computed as follows:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}} \tag{5}$$

where $a_i = \sum_j n_{ij}$, $b_j = \sum_i n_{ij}$ *and* $n_{i,j}$ is defined above in Section II-A.

D. Variation of Information

Meila et. al proposed a new measure based on the entropy between two partitions, called *Variation of Information* (**VI**) [18]. Given two partitions C and C' on the element of X, the **VI** distance measure between and C and C', $\mathbf{VI}(C, C')$, is defined as follow:

[box=]equation **VI**(C, C') = H(C) + H(C') - 2I(C, C') Where:

$H(C)$ represents the marginal entropy, defined as follow :

$$H(C) = \sum_{i=1}^{k} P(i) log_2 P(i)$$

with $P(i)$ being the probability that an element belongs to class C_i, calculated as:

$$P(i) = \frac{|C_i|}{n}$$

The term $I(C, C')$ represents the mutual information between C and C' defined as follows:

$$I(C, C') = H(C) + H(C') - H(C \times C') \tag{6}$$

Therefore, the mutual information, $I(C, C')$, can rewritten as follows:

$$I(C, C') = \sum_{i=1}^{k}\sum_{j=1}^{l} P(i, j) log_2 \frac{P(i, j)}{P(i)P(j)} \tag{7}$$

Where

$$P(i, j) = \frac{\left|C_i \cap C'_j\right|}{n}$$

is the probability that an element belongs to class C_i in C and to class C'_j in C' is

E. Normalised Variation of Information (NVI)

The *Normalised Variation of Information* (**NVI**) is a distance measure defined by Free et al. in [12]. Given two partitions C and C', The **NVI** distance between C and C' is defined as follows:

$$\mathbf{NVI}(C, C') = \frac{H(C) + H(C') - 2I(C, C')}{H(C) + H(C')} \tag{8}$$

By defining $Precision$ and $Recall$ as follows :

$$Precision = \frac{I(C, C')}{H(C)} \quad and \quad Recall = \frac{I(C, C')}{H(C')}$$

$F_{Measure}$, defined to be the evenly weighted, harmonic mean of the $Precision$ and the $Recall$ can be formulated as follows:

$$\begin{aligned} F_{Measure}(C, C') &= \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \\ &= 2\frac{I(C, C')}{H(C) + H(C')} \\ &= 1 - 2\mathbf{NVI}(C, C') \end{aligned} \tag{9}$$

Therefore, $\mathbf{NVI}(C, C')$ can be written in terms the $F_{measure}(C, C')$ as follows:

$$\mathbf{NVI}(C, C') = \frac{1 - F_{Measure}(C, C')}{2} \tag{10}$$

III. Framework of the study

We expect a good distance to be less sensitive to small perturbations, more sensitive to large variations and less impacted by the scale of data. Which can be expressed by the following three intuitions

- Does the distance capture the variation in case of randomly generated partitions? This intuition tries to capture either the distance measure uses its range or it takes only few values
- Is the distance sensitive to small perturbations? This intuition tries to capture the stability of the measure to small perturbation within the data
- Is the distance independent of the scale of the dataset? This intuition tries to measure the ability of the distance measure to capture the "shape" of the data independently of the size of the dataset

One of the main objectives of this paper is to investigate how the distance and the similarity measures discussed above perform in capturing these three intuitions.

Three experiments are conducted, to compare the behaviour of several partition indices and distance measures, **RI** [16] , **ARI** [17], **VI** [18] and **NVI** [12], with respect to the

predefined intuitions above.

In all the three experiments, the mean value of each of the considered measures in this paper as well as its coefficient of variation (standard deviation over the mean) are calculated. These three experiments are described in details in Section IV.

IV. Experimental comparative study

In this section a detailed explanation of each of the three experiments discussed in this work is given. The result for each experiment are discussed and the performance of the algorithms: **RI** [16] , **ARI** [17], **VI** [18] and **NVI** [12] are compared compared against each other.

A. First Experiment - Variation within Randomly Generated Partitions

The purpose of this first experiment is to compare **RI**, **ARI**, **VI** and **NVI** in their ability to capture the data dispersion in the case of randomly generated partitions. The experiment was conducted as follows :

- First, two partitions of 10 000 elements were generated, over the same dataset, with the same predefined number of clusters.
- The distance between the two partitions was then calculated using each the distance measure discussed in this paper (**RI**, **ARI**, **VI** and **NVI**) .
- This process was repeated 30 times. The mean, the standard deviation and the standard deviation over the mean were then calculated for each distance measure.
- This experiment was repeated for each of the following number of clusters: $10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500,$ 1000,1500, 2000 and 3000

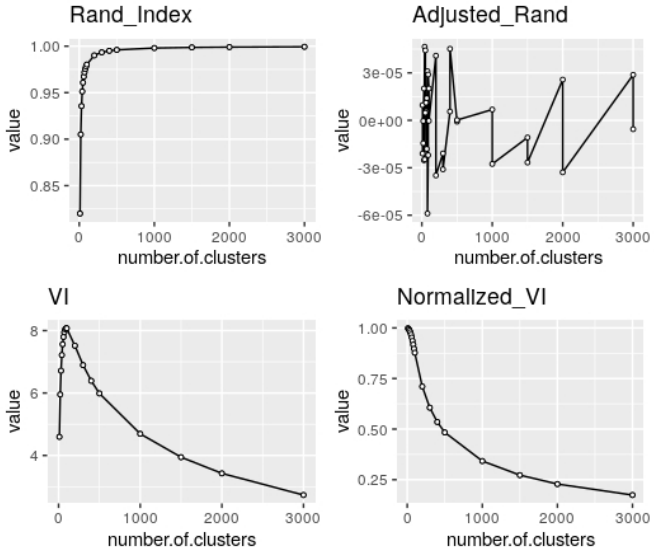

Fig. 1. The Mean value of each distance for each number of clusters.

Figure 1 shows the calculated mean value of each distance measure with respect to the number of clusters. This figure shows that **RI**, which is a similarity index, approaches the value 1 (its upper bound) as the number of clusters increases. This means that the partitions tend to be considered as similar as the number of clusters increases substantially, which is undesirable for two randomly generated partitions. This unexpected behaviour can be explained in Equation 11 where the value n_{00} (number of pairs different clusters and C and C') increases and becomes a dominant term:

$$\mathbf{RI}(C, C') = \frac{n_{00} + n_{11}}{n_{11} + n_{10} + n_{01} + n_{00}} \rightarrow \frac{n_{00}}{n_{00}} \rightarrow 1 \quad (11)$$

Unlike in the case of *Rand Index* (**RI**) where the index approaches the value 1 for large number of clusters, this undesirable behaviour, has been corrected in the case of *Adjusted Rand Index* (**ARI**). However, as it is shown in Figure 1 the **ARI** takes negative values, which is not desirable for a similarity measure. Figure 1 also shows that in the case of *Variation of Information* (**VI**) , the mean value increases with the number of clusters and starts decreasing when the number clusters exceeds 100. However, this initial increase observed in the case of **VI** has disappeared when *Normalised Variation of Information* (**NVI**) is used instead.

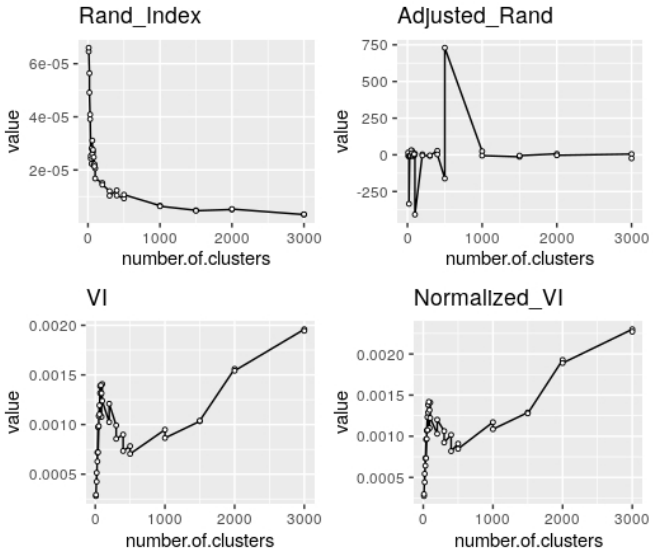

Fig. 2. The SD/MEAN value of each distance for each number of clusters

Figure 2 shows how the coefficient of variance, standard deviation over the mean, varies with respect to the number of clusters. The Figure shows that the **ARI** captures better the variation in randomly generated clusters. Whereas, the remaining measures, **VI**, **NVI**, and **RI** are less sensitive to this variation.

B. Second Experiment -Robustness to Small Perturbations

The main purpose of this experiment is to measure how sensitive is a distance to small perturbations. We expect a good distance to be less sensitive to small perturbations. To achieve small perturbations, the experiment was conducted as follow :

- First, one partition of 10 000 elements is generated,
- Then, a new partition is generated from the first partition by re-affecting randomly one data point to a new class.
- The distance between the two partitions is then calculated using each distance measure considered in this paper: **RI**, **ARI**, **VI** and **NVI**.
- This process is repeated 30 times. The mean, the standard deviation over the mean is then calculated for each distance measure.

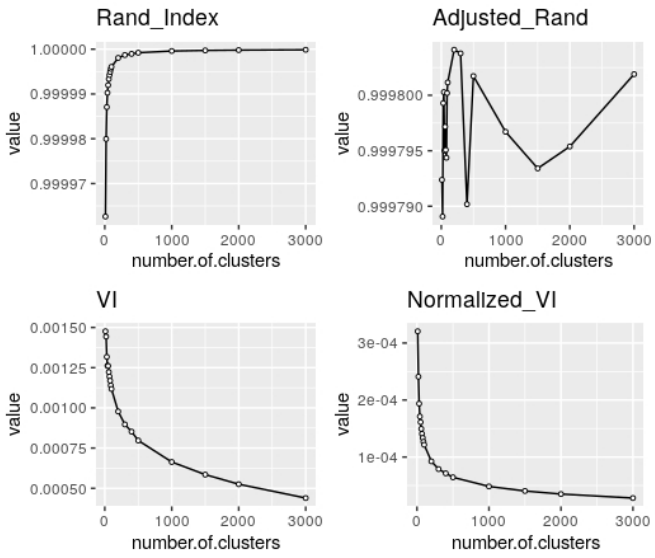

Fig. 3. The Mean value of different measures with respect to the number of clusters.

The Figure 3 shows that all the measures discussed in the paper succeeded to capture the closeness of partitions: The mean value of similarity index for both **RI** and **ARI** approach the value 1 whereas, the remaining distance measures: **NVI**, **VI** approach the value 0, which is expected for a distance measure in the case of close partitions.

Figure 4 shows the variation of the coefficient of variance, standard deviation over the mean, with respect to the number of clusters. This figures shows also that, compared to the other studied measures, the **ARI** is less sensible to small perturbations. Whereas, the remaining distance measures , **VI** and **NVI** are more sensitive to small perturbations within the generated partitions.

C. Third Experiment - Sensitivity to the size of the dataset

This third experiment focuses on comparing the sensitivity of the **VI**, **RI**, **ARI** and **NVI**, when the size of the dataset is doubled. In other words, given a dataset, if we consider a pair of partitions and measure their distances and replace in the underlying dataset each element by 2 elements, will the distance between the new generated pair of partitions be different ?

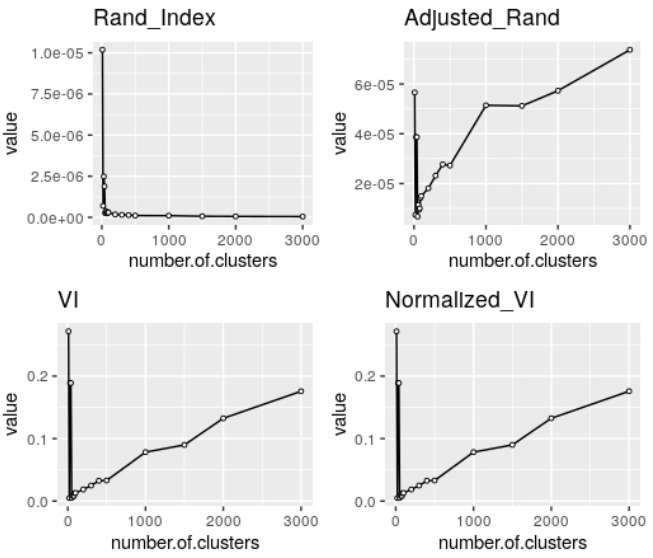

Fig. 4. The SD/MEAN values for different measure with respect to different number of clusters.

This experiment is conducted as follows:

- First, 2 partitions of 10000 elements, are generated randomly with the same number of clusters
- The distance between the 2 partitions are calculated, for all the distance measures considered in this paper
- Then 2 new partitions are generated from the 2 previous partitions by duplicating each element in the dataset
- The distance between the 2 new partitions is then calculated, for all the distance measures considered in this paper
- Then, the mean, standard deviation and standard deviation over the mean of the measured distances are calculated
- The following number of clusters were considered: $10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, 1500, 2000, 3000$

We then generate 2 new partitions from the 2 previous partitions by duplicating each element in the dataset and recalculating their distances, using the four partition measures discussed in this paper. The mean, standard deviation and standard deviation over the mean of the measured distances are then calculated. The following number of clusters were considered:
$10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500, 1000, 1500, 2000$ and 3000

In Figure 5, the standard deviation over the mean, with respect the number of clusters for each distance measure is shown.

The Figure shows that the **VI** and **NVI** are less sensible to the scale of the dataset, whereas **ARI**, and **RI** are more sensitive.

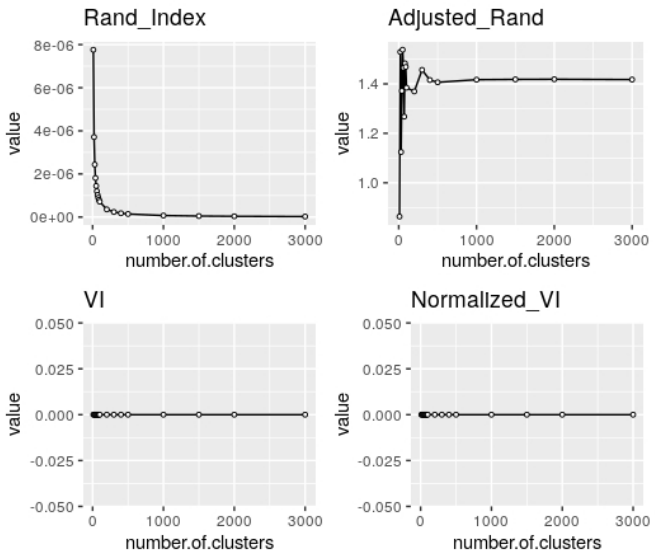

Fig. 5. Comparison between the SD/mean values

V. Conclusion and Future Work

A. Conclusion

In this paper a comparison between **RI**, **ARI**, **VI** and **NVI**, has been conducted via an empirical study given in Section IV. This empirical study consists of three experiments and the results of these experiments are summarised as follows:

- The first experiment in Section IV-A shows that **ARI** captures better the variation in the case of randomly generated partitions.
- The second experiment given in Section IV-B demonstrates that **ARI** is less sensitive to small perturbations.
- Finally, the third and final experiment discussed in Section IV-C shows the **VI** and **NVI** are less sensitive to the scale of the data.

In conclusion, the results of these experiments show that **ARI**, performs well in capturing the intuitions defined earlier.

VI. Limitations and future Work

In one hand, the current work have been conducted under several assumptions such as predefined dataset size of 10 000 points. Also, the distance and similarity measures are calculated between partitions having the same number of clusters i.e. $10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 200, 300, 400, 500$, 1000, $1500, 2000$ and 3000 partitions. In addition, the experiments realised were focused on measuring the ability of the distance and similarity measures to capture three main intuitions regarding the variation, the sensitivity and the scalability of the dataset.

Future work can evaluate other intuitions that a "good" distances can have, such as independence of the distance measure from the dataset size, from the number of clusters and from the cluster sizes.

On the other hand, this paper focus was on measuring the mean, and the standard deviation over the mean. Future work can be extended to capture other properties of the distance measures by examining other statistical such as kurtosis and and skewness of the distance measures distributions.

References

[1] A K Jain and M N Murty and P. J. Flynn, *Data Clustering: A Review*, 1999

[2] E. B. Fowlkes and C. L. Mallows, A Method for Comparing Two Hierarchical Clusterings, *Journal of the American Statistical Association*,1983

[3] Jolliffe, Ian T. and Morgan, Byron J. T, Comments on A method for comparing two hierarchical clusterings, *Journal of the American Statistical Association*,1983

[4] Santos, Jorge M. and Embrechts, Mark, On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification, *Artificial Neural Networks – ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part II, Springer Berlin Heidelberg* ,2009

[5] Denœud, Lucile and Guenoche, Alain, Comparison of Distance Indices Between Partitions, *Springer Berlin Heidelberg*,2006

[6] Daniel Cosmin Porumbel and Jin Kao Hao and Pascale Kuntz, An efficient algorithm for computing the distance between close partitions, *Discrete Applied Mathematics* ,2011

[7] Morlini, I. and Zani, S., An Overall Index for Comparing Hierarchical Clusterings, *Challenges at the Interface of Data Analysis, Computer Science, and Optimization: Proceedings of the 34th Annual Conference of the Gesellschaft für Klassifikation e. V., Karlsruhe, July 21 - 23, 2010*,2012

[8] Charon, Irene and Denoeud, Lucile and Guenoche, Alain and Hudry, Olivier, Maximum Transfer Distance Between Partitions, *Journal of Classification*,2006

[9] Gardner, Andrew and Kanno, Jinko and Duncan, Christian A. and Selmic, Rastko, Measuring Distance Between Unordered Sets of Different Sizes, *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*,2014

[10] Boris Mirkin, Eleven Ways to Look at the Chi-Squared Coefficient for Contingency Tables, *The American Statistician*,2001

[11] Mohamed Bouguessa and Shengrui Wang and Haojun Sun, An objective approach to cluster validation, *Pattern recognition letters*,2006

[12] Ana L. N. Fred and Anil K. Jain, Robust Data Clustering, 2003

[13] Strehl, Alexander and Ghosh, Joydeep, Cluster Ensembles — a Knowledge Reuse Framework for Combining Multiple Partitions, *J. Mach. Learn. Res.*,2003

[14] J. M. Santos and S. Ramos, Using a clustering similarity measure for feature selection in high dimensional data sets, *2010 10th International Conference on Intelligent Systems Design and Applications*,2010

[15] Silke Wagner and Dorothea Wagner, Comparing Clusterings- An Overview, 2007

[16] Rand, William M., Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*,1971

[17] Hubert, L. and Arabie, P., Comparing partitions, *Journal of classification*,1985

[18] Marina Meila, Comparing Clusterings, 2002