Towards Cluster Validity Index Evaluation and Selection

Andrey Filchenkov, Sergey Muravyov, Vladimir Parfenov
ITMO University
St. Petersburg, Russia
{afilchenkov,smuravyov}@corp.ifmo.ru, parfenov@mail.ifmo.ru

***Abstract*—In this work, we address the hard clustering problem. We study how well clustering algorithm efficacy measures (clustering validity indices) can reflect the clustering quality. We use assessors' estimations for cluster partition adequacy as the ground truth and explain, why this is the only measure that can be used in this quality. We compare different clustering validity indices and show that none of them can be the universal, reflecting quality for each cluster partition. To do so, we introduce four quality measures for CVI evaluation. Also, we suggest an approach for the best CVI prediction for a given dataset based on meta-learning.**

I. Introduction

Cluster analysis is the art of finding groups in data, these groups are called clusters. Cluster analysis is used in many domains, such as geography, medicine, chemistry and many others [1]. In this paper, we address the problem of *hard clustering*, in which it is assumed that each point belongs to a single cluster [2].

Clustering problem is initially formulated in general terms. Many mathematical formalizations were suggested to describe it. It was recognized quite a long time ago that the selection of a certain formalism to estimate the clustering quality will strongly effect both the algorithm that should be chosen and result claimed to be the best [3], and that it is hard to compare different formalisms [4]. Kleinberg formulated three essential properties of clustering algorithm and proved that there is no way to build an algorithm that fits all these properties simultaneously [5].

The problem of algorithm performance evaluation is fundamental in any computer science domains. But this problem becomes really sharp when one tries to solve clustering algorithm selection problem, which is to predict the best clustering algorithms (or their ordering) for a never seen dataset. It can be solved with meta-learning [6], [7] that reduces this problem to a supervised learning problem. One of the key elements in this reduction is a quality performance measure that is essential for crafting labels for the dataset.

There have been several suggestions made on how to measure a similarity between two cluster partitions. This measure can be used to compare how well a clustering algorithm performs on a dataset. These measures usually depend on the type of criterion that is considered in assessing the quality of a clustering algorithm. Nowadays, there are two classes of clustering metrics: external and internal [8].

Evaluation with *external metrics* is based on data that was not used for clustering, such as known class labels and external benchmarks. We must point out that this approach has theoretical restrictions. These restrictions include not only the additional data required, but also the fact that a clustering problem usually allows to obtain different adequate partitions into clusters, but this approach is blind for all of them except the partition formed by the labels. This practice may be considered as at least questionable; for detailed discussion see [9], [10]. Evaluation with *internal metrics* is based only on data partition. These metrics usually assign the best score to an algorithm that returns partition with high similarity within a cluster and low similarity between clusters. There are plenty of different internal metrics nowadays, and they appear at a very high rate. We will refer to these metrics as *cluster validity indices* (CVI). There are several works that are devoted to the comparison of different CVIs behaviour [11], [12], but all of them state different CVI to be the best on certain types datasets that have structure of well-separated hyper spheres. Thus we can conclude that there is no perfect CVI to this moment.

This paper addresses the problem of cluster validity index evaluation. The contribution of this paper includes: a) a framework for evaluation CVIs based on human evaluations; b) four quality measures for CVI evaluation; c) comparison of 19 well-known CVIs with respect to the proposed measures; d) approach for the best CVI prediction for a given dataset based on meta-learning.

The structure of this paper is following. In Section II, we provide a theoretical foundation for understanding how to choose an algorithm performance measure and how to evaluate quality of choice, explaining why it can be done only on the basis of human estimations. This is basement for the framework for clustering algorithm performance measures comparison, described in III. Then we compare 19 CVIs, the results of that are presented in Section IV. In Section V, we discuss if meta-learning can be applied to the problem of algorithm performance measure selection. Section V concludes the paper with a summary of the work done and propositions on future work.

II. Fundamental assumptions for CVIs evaluation

A. Clustering validity indices validation

First of all, we describe validation techniques that are used to justify a novel CVI. We have found four main strategies to validate a CVI quality and applicability:

1) Visual-based comparison. As an example, see paper [13].

ISBN 978-952-68397-8-3 (PDF)

2) Comparison with known labels. This validation technique is very popular. Examples of papers which utilize it are [14], [15], [16], [17], [18], [19].
3) Purely theoretical comparison that is based on studying CVI properties. Examples of papers that utilize this methodology are [20], [21], [22].
4) Comparison based on stability, robustness of structure, or other desired properties, estimation of which requires a transformation (usually, subsampling) of initial dataset. As examples see [23], [24].

It is worth to note that datasets, which are used within the first technique, usually have a single obvious partition. As a result, application of this technique to these datasets cannot be distinguished from the application of the second technique. They share the same idea with external metrics: comparison with a single and a priori known "ideal" partition. Despite being cheap to implement, this technique is a subject of criticism similar to the one of external metrics. As we pointed out in the Introduction, the key problem is that this ideal partition is expected to be single. Also class labels may not create a good partition in any term. This is why such intention to use a clustering algorithm as classifiers is methodologically inconsistent.

Comparison of stability is based on an assumption that clusters should be insensitive to any data change. However, this is another restriction to application of clustering algorithms. It cannot be considered as a universal criterion.

As it is stated in the classical book [25], "Understanding our world requires conceptualizing the similarities and differences between the entities that compose it". Thus, clustering is a very natural task that is being performed all the time in our brain during cognition. Therefore, we can assume that a human can properly estimate the quality of a clustering partition. Moreover, as we have shown, it is the only way of obtaining ground truth for partition quality estimation.

Now we get to the point that human estimation is the only credible point to estimate clustering adequacy. All the existing techniques either can be reduced to the human evaluation, or are methodologically unjustified. In [26], authors showed that there is no difference in partition evaluations made by clustering experts and non-experts. This is another justification for the assumption we make about human-based ground truth. In the next subsection we formalize this assumption.

It is clear that humans cannot estimate partitions in a high-dimensional spaces, in which objects are represented as points. In these cases, hoverer, no appropriate solutions can be found. Applicability of CVIs in high-dimensional case can be only evaluated either as predicting known labels, or with usage of external quality measure. As we described above, these approaches are methodologically inconsistent. To handle non-2D datasets, we use multidimensional scaling, that results in serious restrictions on the method applicability. However, multidimensional data visualization for human estimation is a hot topic and several studies are being conducted on how to make human estimations applicable for higher dimensions (as an example, see [27], where techniques for visual assessment of subspace clusters are presented).

B. Ground truth for partitions

The space of clustering problems is simply the universal set of point sets; we will denote it as $\mathscr{C}$. Note that we do not specify the space, in which the points are located. It can be a vector or a metric space, or even a latent space specified by a distance or another similarity measure. From this point of view, each clustering problem is a set of points with no other assumptions. This understanding is important, due to the contemporary trend is making the assumption that different clusters are generated by different processes. This assumption restricts the area of clustering algorithms application.

For each clustering problem (set of points) $c \in \mathscr{C}$, $\mathrm{Partition}(c)$ denotes the set of possible partitions of c, and $\mathbf{Partition}$ denotes all partitions of all $c \in \mathscr{C}$. CVI is a function defined on $\mathbf{Partition}$. Without loss of generality, we assume that codomain of all CVIs is $[0; 1]$, where 1 corresponds to the best partition.

We proclaim the existence of the human-generated ground truth representing quality of each cluster partition. However, one can suggest several ways of how this ground truth can be formalized. The natural way to do so is to assume that a strict weak ordering exists of each $\mathrm{Partition}(c)$. This means that any two elements of $\mathrm{Partition}(c)$ either can be compared ("which is better?") or are incomparable, and incomparability is transitive relationship. With this assumption, a CVI can be understood as a ranker predicting the ordering of partitions.

We additionally assume that there is a clear mapping from $\mathrm{Partition}(c)$ to a binary scale to which we will refer as "adequate" and "inadequate". In other word, each partition can be related to one of these categories describing its quality. With this assumption, a CVI can be understood as a binary classifier predicting a partition quality. These two assumptions are independent in general. However, we expect that "adequate" partitions are never worse than "inadequate".

We formulate the assumptions above more precisely. Choose and fix a clustering problem $c \in \mathscr{C}$. Let H_r be a function that represents weak strict ordering:

$$H_r(c) : \mathscr{C} \to \mathrm{Partition}(c) \times \mathrm{Partition}(c)$$

and H_b be a function that represents a binary estimation scale:

$$H_b : \mathrm{Partition}(c) \to \{0; 1\}.$$

Each CVI can be understood as an approximation of H_r or H_b. In the first case the quality of CVI can be measured as

$$\mathbf{E}\,\mathrm{Sim}(\underset{CVI}{\mathrm{Rank}}\,\mathrm{Partition}(\pi), H_r(\pi)),$$

where $\mathbf{E}$ is expectation, π is a random variable representing a point set from $\mathscr{C}$ and Sim is a similarity measure between two weak strict orderings. In the second case the quality of CVI can be measured as

$$\mathbf{E}\,L(CVI(\gamma), H_b(\gamma)),$$

where γ is a random variable representing obtaining a partition from $\mathbf{Partition}$ and L is a loss function (as for binary classification).

III. Framework for CVIs evaluation

A. Procedure for one assessor

Assume that we have selected several clustering problems. First, we describe a procedure for evaluating both the strict weak order of partitions and their adequacy with a single respondent (assessor) for each selected clustering problem.

The perfect way to elicit this information from the assessor is to ask her/him to evaluate each possible partition for a given c. However, the number of possible partitions growth unmanageably fast with the increase of the number of points in the dataset (see Hardy—Ramanujan—Rademacher formula [28]) thus making it practically impossible. This is why we need to consider only a subset of partitions. Another important thing is that elements of these subsets should be better than average, because only a few of partitions for given set have high scores.

A strict way to evaluate the partitions by assessors is to make a partial order by comparing pairs of partitions with each other. However, this approach is very time-consuming and nearly impossible to perform. To overcome this, we use the fact that there is a score function (sometimes called utility function) that can be put into correspondence to any weak strict order. Exploiting this fact, we can ask the assessor to mark each partition from the selected subset with a number. If the assessor found that this partition is adequate, the assigned value should positive, otherwise it should be negative. The higher the value that the assessor assigns, the more adequate is the partition. Assessors should be able to assign the same values to different partitions, it means that the assessor cannot find the difference between the partitions quality. As a result, for each dataset all the partitions receive marks from each assessor. These values are used to recover strict weak ordering, corresponding to assessors consideration.

B. Procedure for several assessors

Assume that we have selected several clustering problems, and for each clustering problem we selected a subset of partitions. Now we describe a procedure for working with several assessors. Using the procedure described in the previous subsection, we can recover evaluation for each of the assessors for each of the clustering problems.

We suggest four criteria to estimate a CVI quality: binarized adequacy (BA), weighted adequacy (WA), adequacy of the best (AB), and aggregated ranking (AR).

1) Binarized adequacy: To estimate the binarized adequacy of CVI for a given clustering problem c, we performed the following steps. Each partition received assessors estimations, positive or negative. If more than a half of assessors found this partition adequate, then it was marked as adequate (+). If less than a half of assessors found this partition adequate, then it was marked as inadequate (-). Then for each CVI, we ordered assigned marks with respect to the CVIs estimations of these partitions, thus, each CVI was characterized by a permutation on marks r_{CVI}. In this context, the ideal ordering r^{BA}_{best} is the one, in which adequate partitions have the highest values and inadequate partitions have the lowest values: $(+\ldots+-\ldots-)$. The worst permutation r^{BA}_{worst} is the opposite one: $(-\ldots-+\ldots+)$. To estimate CVI average adequacy, we estimated distance between the corresponding permutation and the ideal one. To normalize this value, we divided it to the distance between the ideal and the worst orderings.

Formally, BA is calculated in the following way:

$$R^{BA} = \frac{\rho_{K\tau}(r^{BA}_{best}, r_{CVI})}{\rho_{K\tau}(r^{BA}_{best}, r^{BA}_{worst})},$$

where function $\rho_{K\tau}$ denotes modified Kendall tau distance [29].

2) Weighed adequacy: In order to obtain weighed adequacy, we simply follow the previous approach with the only exception: instead of assigning adequate (+) or inadequate (-) mark to each partition, we assign to it the exact number of assessors, who thought it be adequate. Then a value w in range $0..n$, is assigned each partition, where n is the number of assessors. The best ranking r^{WA}_{best} is now $(w_1, \ldots, w_m)$, such that $w_1 \geq \ldots \geq w_m$, and the worst ranking r^{WA}_{worst} is now $(w_1, \ldots, w_m)$, such that $w_1 \leq \ldots \leq w_m$.

Formally, WA is calculated in the same way as BA:

$$R^{WA} = \frac{\rho_{K\tau}(r^{WA}_{best}, r_{CVI})}{\rho_{K\tau}(r^{WA}_{best}, r^{WA}_{worst})}.$$

3) Adequacy of the best: In order to obtain adequacy of the best, we simply chose the adequacy mark, which was assigned to the partition having the highest value of CVI being estimated. We chose both BA and WA mark.

4) Aggregated ranking: The last measure represents how many orderings produced by the assessors and by each CVI differ. Since a lot of these orders are weak, we apply weak order aggregation algorithm and distance measure [30]. As in the first case, we normalize the distance between ordering by dividing it to the distance between assessors order and an opposite order.(An opposite order is a strong order, which is produced by inverting the original weak order and randomly choosing a consistent strong order.)

IV. Experimental evaluation

A. List of CVIs in comparison

We examine 19 of the most popular CVIs to find out if any of them matches the real quality of resulting clusters. Most of the indices estimate the cluster cohesion (within- or intra-variance) and the cluster separation (between- or inter-variance) and combine them to compute a quality measure. The combination is performed by division (ratio-type indices) or summarization (summation-type indices). More detailed description of each metric can be found in [31]

1) Dunn index (D) is a ratio-type index where the cohesion is estimated by the nearest neighbor distance and the separation by the maximum cluster diameter.
2) Davies-Bouldin index (DB) is an index which estimates the cohesion based on the distance from the points in a cluster to its centroid and the separation based on the distance between centroids.
3) Silhouette index (Sil) is a normalized summation-type index. The cohesion is measured based on the distance between all the points in the same cluster

and the separation is based on the nearest neighbor distance.

4) Calinski–Harabasz (CH) is a ratio-type index where the cohesion is estimated based on the distances from the points in a cluster to its centroid. The separation is based on the distance from the cluster centroids to the global centroid, which is denoted as the mean vector of the whole dataset.
5) CS index (CS) was proposed in the image compression community, but can be extended to any other environment. It is a ratio-type index that estimates the cohesion by the cluster diameters and the separation by the nearest neighbor distance.
6) C-Index (CI) is a type of normalized cohesion estimator.
7) Davies-Bouldin* index (DB*) is a variation of the classical Davies-Bouldin index. Separation between clusters is minimal throughout all the clusters comparing to original DB index.
8) Sym-index (Symm) is an adaptation of the I index based on the Point Symmetry-Distance.
9) S_Dbw index (SDbw) is a ratio-type index that has a complex formulation based on the Euclidean norm, the standard deviation of a set of objects, and the standard deviation of a partition.
10) Score function (SF) is a summation-type index in which the separation is measured based on the distance from the cluster centroids to the global centroid, and the cohesion is based on the distance from the points in a cluster to its centroid.
11) COP-index (COP) is a ratio-type index in which the cohesion is estimated from the distance from the points in a cluster to its centroid and the separation is based on the furthest neighbor distance.
12) SV-Index (SV) is a ratio-type index, which is one of the most recent CVIs compared in this work. It estimates the separation by the nearest neighbor distance and the cohesion is based on the distance from the border points in a cluster to its centroid.
13) OS-Index (OS) is another ratio-type index very similar to the SV-Index, where a more complex separation estimator is used.
14) Generalized Dunn indices gD31, gD41, gD51, gD33, gD43, gD53. All the variations are different combinations of three variants of the separation estimator and two variations of the cohesion estimator.

B. Experiment setup

Real datasets can have very different topologies of clusters, but usually it is very problematic to find pure examples of a topology of a certain type. This is why we preferred to use not the real-world datasets, but synthetic ones that were generated manually. The practice of using synthetic dataset is common is cluster analysis, see [32] for discussion. We used 41 different datasets of different forms. In order to simplify the work of assessors, only two-dimensional datasets were taken into account.

In order to get a subset of partitions, we applied standard clustering algorithms. We could also synthesize artificial partitions, but it seems that random partitions are very unlikely to achieve high results, thus we decided to use tools, which were designed to achieve this goal — existing clustering algorithms. We used six clustering algorithms implemented in WEKA library [33]. These algorithms are:

- k-Means [34];
- X-Means [35];
- EM [36];
- DBSCAN [37];
- FarthestFirst [38];
- Hierarchical [39].

k-Means, EM, X-Means and FarthestFirst were exploited several times with different seeds, because of their probabilistic nature. Euclidean distance was taken dissimilarity measure between instances in the dataset. k-Means was run with maximum number of iterations equal to 500 and preset number of clusters is 2. EM algorithm was exploited with parameters: the number of clusters equal 2, the maximum number of iterations equal to 100, and minimum standard deviation 10^-6, DBSCAN was run with parameters: epsilon equal to 0.1 and the minimum number of elements in cluster equal to 6. In FarthestFirst algorithm, the number of clusters equaled 2. Hierarchical algorithm was single-linkage and was run with the number of clusters equal to 2. X-means had the following parameters: the maximum number of iterations to perform was equal to 1, the maximum number of iterations to perform in k-Means was equal to 1000, the maximum number of iterations k-Means that were performed on the child centers was equal to 1000, the maximum number of clusters was equal to 2.

Thus, for each dataset we obtain six partitions. All the results were structured into a table. After that 5 independent assessors' evaluations were made. Assessors were given pictures, one for each dataset, and each picture contained initial dataset, and all the cluster partitions produced by the described algorithms. Examples of such pictures and datasets mentioned above are presented in Fig. 1 and Fig. 2 (only 3 of 6 used partitions are presented). The full set of images can be found online https://www.dropbox.com/sh/1qn6ydqtd452y0h/AAA7r6wE7BdVPb_qqJp2K9W_a?dl=0. The datasets can be found online https://www.dropbox.com/sh/7ybw2s3ljskdxbl/AADJ5luBcYpItAJ-rwyIv_vNa?dl=0.

C. Comparison results

The results are summarized into three tables. Table I shows BA, Table II shows AR, Table III shows WA and Table IV shows AB. For average adequacy and average rank we used the following strategy: for each cluster partition we took three the best weight values among all the indices and marked them. After that we counted the relation of the marked values to the total number of cluster partition results. These relations are presented in Table I and Table II. In this table partition is called the best, if more than a half of the assessors think the result is adequate, and good, if only a half of the assessors think the result is adequate)

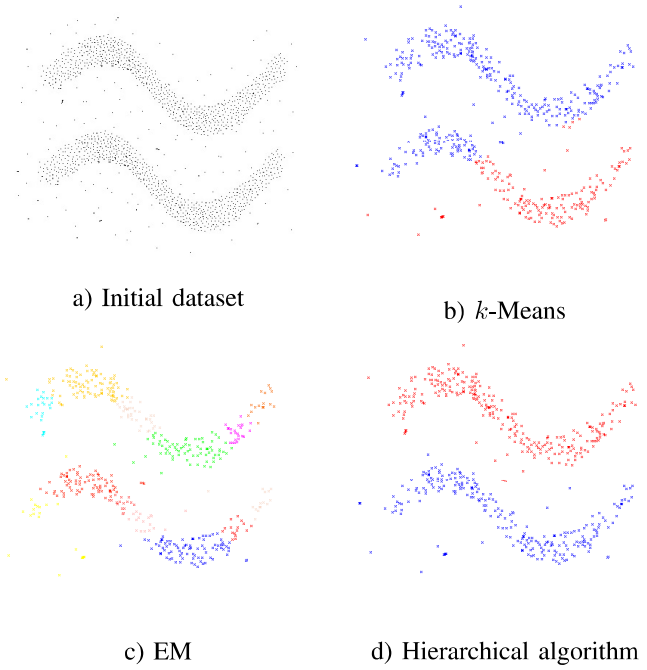

Fig. 1. Depicted dataset that looks like curved stripes and different partitions generated by different clustering algorithms: a) initial dataset; b) k-Means partition; c) EM partition; d) hierarchical algorithm partition.

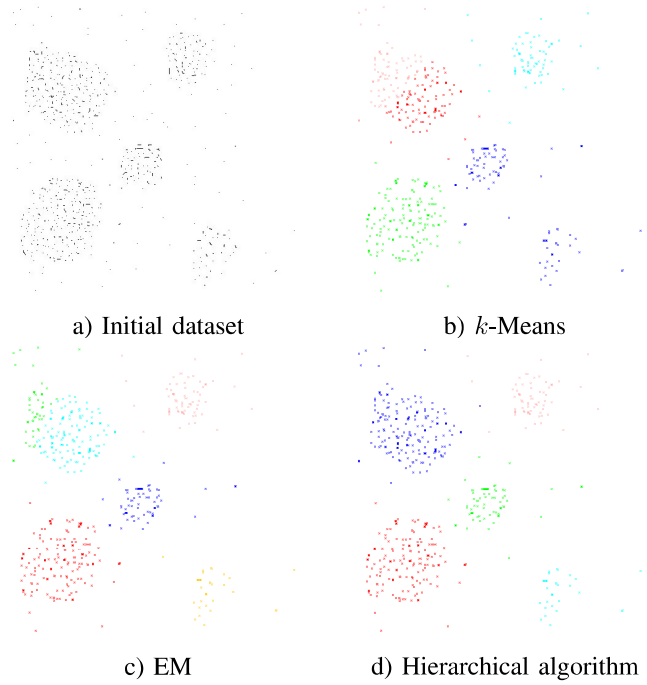

Fig. 2. Depicted dataset that contains five clusters with noise and different partitions generated by different clustering algorithms: a) initial dataset; b) k-Means partition; c) EM partition; d) hierarchical algorithm partition.

After that we run each of the 19 CVIs and evaluated corresponding rankings. Then for each CVI and each dataset we evaluated each of the four measures, described in the previous Section.

TABLE I. SUMMARIZATION OF BINARIZED ADEQUACY OF CVIS.

INDEX	ADEQUACY	INDEX	ADEQUACY
DB	0.151	SYMM	0.333
D	0.515	CI	0.121
SIL	0.393	DB*	0.121
CH	0.242	GD31	0.333
SDBW	0.454	GD41	0.424
SF	0.393	GD51	0.393
CS	0.303	GD33	0.242
COP	0.575	GD43	0.272
SV	0.272	GD53	0.333
OS	0.575	**threshold**	**0.7**

TABLE II. SUMMARIZATION FOR AGGREGATED RANK FOR EVERY CVI

INDEX	RANK	INDEX	RANK
DB	0.390	SYMM	0.292
D	0.243	CI	0.243
SIL	0.170	DB*	0.317
CH	0.073	GD31	0.073
SDBW	0.243	GD41	0.146
SF	0.146	GD51	0.195
CS	0.146	GD33	0.024
COP	0.414	GD43	0.024
SV	0.414	GD53	0.073
OS	0.073	**threshold**	**0.7**

We assume that an index is adequate and applicable when the value of the index presented in Table I, Table II and Table III is not less than 0.7 and the number of adequate good results in Table IV is more than 70% from the number of clustering result.

None of the metrics fits the requirement, described above. Thus, we have shown that there is no perfectly applicable, universal metric. However, if we look at the results more precisely, we can figure out that there is an index metric, standing out of all others, it is COP. This index can be used under some conditions for some tasks.

V. META-LEARNING FOR GROUND TRUTH PREDICTION

The suggested approach has an obvious disadvantage, it is expensive due to the each new dataset has to be evaluated by assessors. This is a typical problem in clustering validation domain, which gave popularity to the label-based evaluation.

In this subsection we describe a way how the suggested approach can be improved in a way to relax or even eliminate this problem based on meta-learning. Meta-learning is a field of machine learning which is focusing on predicting algorithm performance for a given problem. This prediction is based on known algorithm performance on similar problems, thus, meta-learning is closely related to transfer learning.

Let $\mathscr{P}$ denote the problem space (domain), $\mathscr{A}$ denote the algorithm space related to this domain, and Q denote a performance measure for algorithms from $\mathscr{A}$ (more precisely, $\mathscr{A} = \mathscr{A}_{\mathscr{P}}$ and $Q = Q_{\mathscr{P}}$), because they are domain-specific).

TABLE III. SUMMARIZATION FOR WEIGHTED ADEQUACY FOR EVERY CVI

INDEX	ADEQUACY	INDEX	ADEQUACY
DB	0.121	SYMM	0.243
D	0.317	CI	0.073
SIL	0.121	DB*	0.000
CH	0.097	GD31	0.097
SDBW	0.195	GD41	0.097
SF	0.170	GD51	0.121
CS	0.146	GD33	0.073
COP	0.414	GD43	0.024
SV	0.195	GD53	0.024
OS	0.317	**threshold**	**0.7**

TABLE IV. ADEQUACY OF THE CLUSTER PARTITIONS, CLAIMED TO BE THE BEST AND GOOD BY CVIS.

INDEX	# OF GOOD	# OF THE BEST
DB	6 (14.6%)	4 (9.7%)
D	13 (31.7%)	4 (9.7%)
SIL	7 (17.0%)	4 (9.7%)
CH	5 (12.1%)	4 (9.7%)
SDBW	9 (21.9%)	6 (14.6%)
SF	7 (17.0%)	4 (9.7%)
CS	5 (12.1%)	6 (14.6%)
COP	14 (34.1%)	6 (14.6%)
SV	7 (17.0%)	6 (14.6%)
OS	14 (34.1%)	5 (12.1%)
SYMM	8 (19.5%)	5 (12.1%)
CI	6 (14.6%)	4 (9.7%)
DB*	6 (14.6%)	4 (9.7%)
GD31	6 (14.6%)	5 (12.1%)
GD41	6 (14.6%)	5 (12.1%)
GD51	5 (12.1%)	6 (14.6%)
GD33	5 (12.1%)	4 (9.7%)
GD43	6 (14.6%)	4 (9.7%)
GD53	4 (9.7%)	5 (12.1%)
threshold	**70.0%**	

We need to assume that $\mathscr{A}$ is finite (This may seem strange. Despite the fact that the number of known algorithm schemes in finite, most of them have real-valued hyperparameters. This can be solved by understanding that hyperparameters are tuned with other algorithms, the number of which is finite. Thus, an algorithm instance may be understood as the last wrapping algorithm which has no hyperparameters, e.g. non-parametric SMA tuning step size for grid search tuning number of neighbors for kNN. However, in practice a finite subset of non-parametric algorithms is chosen as $\mathscr{A}$.)

First, every problem in $\mathscr{P}$ should be characterized by features, which are called meta-features. Let $F = \{f_1, \ldots, f_{|F|}\}$ be a set of meta-features: $f_i : \mathscr{P} \to \mathrm{codomain}(f_i)$. Next, we need to select a problem subset $P_{\text{train}} = \{p_1, \ldots, p_{|P|}\} \subset \mathscr{P}$, which will be used for training a model representing algorithm performance. The next step is evaluation of performance Q of all the algorithms from $\mathscr{A}$ for all the problems from P, obtaining matrix $\mathbf{Q} = (Q(A_i(p_j)))_i^j, A_i \in \mathscr{A}, p_j \in P$. This matrix is used to produce labels y for problems in P :

1) In case of predicting a single best algorithm (classification),

$$y(p_j) = \arg\max_{A_i \in \mathscr{A}} Q(A_i(p_j)).$$

2) In case of predicting algorithms ordering (learning to rank, which is considered as the most reliable and useful statement, see [40]),

$$y(p_j) = \underset{Q(A_i(p_j))}{\mathrm{Rank}} A_i.$$

3) In case of predicting algorithm performance (regression),

$$y(p_j) = \max_{A_i \in \mathscr{A}} Q(A_i(p_j)),$$

or

$$y(p_j) = (Q(A_i(p_j)))_{A_i \in \mathscr{A}}.$$

Finally, we have a training set, which is $\{(F(p_1), y(p_1)), \ldots, (F(p_{|P|}), y(p_{|P|}))\}$, where $F(p)$ stands for vector $(f_1(p), \ldots, f_|F|(p))$. This training set is used to learn a supervised model corresponding to the selected problem statement. Say, M is this model, trained M is an algorithm, which for any $p \in \mathscr{P}$ returns an answer (best algorithm, algorithm ordering or algorithm(s) performance measure value(s)), which can be used to choose the best expected algorithm. Thus, this model with the proper best algorithm selection rule can be considered as a mapping S which is the solution for the algorithm selection problem. Application of the meta-learning approach to answer Rice's question resulted in the detailed modified framework, which is a working scheme of an algorithm recommendation system, presented in Fig. 3.

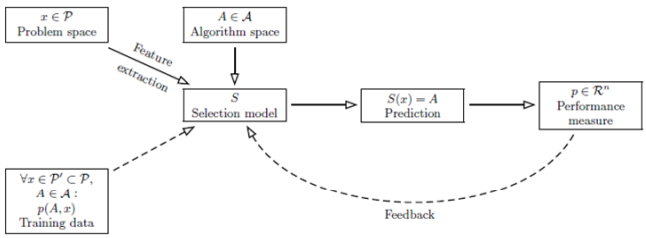

Fig. 3. Rice framework rethought in the meta-learning approach [41]

For a more detailed survey of meta-learning and its application to different domains we encourage the readers to pay their attention to the well-known book [7], and the most recent PhD thesis by Sun [42].

A. Meta-learning for CVI selection

Three steps should be performed to create a meta-learning system: choice of meta-features, choice of training set and choice of performance measure. The last two steps are closely connected in our case, due to assessors' labor is required to work with each dataset.

Selection of meta-feature description seems to be a problem itself. However, as well as data mining algorithms, each CVI is designed under certain assumptions and implements a certain intuition. The transformation of this intuition to numeric features is challenging, but potentially solvable task. We have to notice that this intuition is usually visual. This is why we expect different image characteristics to be fruitful for CVI selection. Also, landmarks can be used that are values of

different CVI on a dataset may be a good predictor for their adequacy.

Talking about selection of training dataset, we must point several questions which must be answered. The first one if we should use only real-world datasets or artificial ones also? On the one hand, meta-learning works under the weak assumption of machine learning [43], which states what learning problems are generated by a process with non-uniform probability over the problem space. On the other hand, clustering is a very popular approach to object space reduction and it is often used as a preprocessing step (which cannot be said about supervised learning algorithms). We suggest that artificial data should be used together with real-world datasets, because doing this, we will gather only additional information about algorithm performance in different cases, some of which can be never met, but we cannot predict in advance, which ones.

Another restriction, which is met for training datasets, is that we need to spend much assessors' labor, and the more datasets we have, the more time should be spent by assessors. To relax these requirements, we can approximate assessors' estimations within a metric space that can be introduced on partitions of a dataset. This space can be introduced in many ways, for instance, by utilizing Rand's index [44]. Then we can approximate H_* on this space with known values of some of its elements.

Another question we need to answer is what to do with datasets which has more than two dimensions. The naïve answer to this question is just to assume that we can expect the same performance on CVI regardless of dataset dimensionality. Thus, the system trained on two-dimensional datasets can be used to predict CVI for other dimensions. A more complicated answer is to use projections of a dataset to several two-dimensional spaces. Despite this approach brings randomness to the estimation process, we expect it to produce more reliable results than the previous one.

VI. Conclusion

In this paper, we have shown that there is no universal clustering validity index existing at the moment. We also state that cluster validity indices should be chosen for problems specifically, and we suggest to apply the meta-learning approach to solve this problem. As the ground truth (and measure of CVI quality), assessors' estimates should be used. After the system is learnt, no more assessors' labor is required, due to the system will be able to predict proper CVI for a given dataset.

This research brings more questions than answers. Questions on how to build a meta-learning system for predicting assessors' evaluation are listed in the previous section. One of the most important questions is the justification of the assumptions we have made about human assessing ability. Another interesting question is how to choose a proper subset of partitions that are representative enough for the entire partition set.

Several questions exist on how to merge assessors' evaluations. Sometimes these evaluations are more consistent, sometimes they are less. This is also related to the fact that some datasets are well-clusterable, and some are not [45]. These considerations should be taken into account during creating a more sensitive evaluation framework.

Besides answering these questions, our future work will be devoted to creating a meta-learning system for predicting human evaluations based on meta-learning. The main problem is feature engineering, however, a lot of other questions arise connected both to theoretical and engineering aspects.

Acknowledgements

Authors would like to thank Georgiy Korneev and Margareta Ackerman and unknown reviewers for useful comments. The research was supported by the Government of the Russian Federation (grant 074-U01) and the Russian Foundation for Basic Research (project no. 16-37-60115).

References

[1] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344.

[2] D. J. Bora, D. Gupta, and A. Kumar, "A comparative study between fuzzy clustering algorithm and hard clustering algorithm," *International Journal of Computer Trends and Technology*, vol. 10, no. 2, pp. 108–113, 2014.

[3] M. R. Anderberg, *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. Academic press, 1973.

[4] R. E. Bonner, "On some clustering techniques," *IBM journal of research and development*, vol. 8, no. 1, pp. 22–32, 1964.

[5] J. Kleinberg, "An impossibility theorem for clustering," *Advances in neural information processing systems*, pp. 463–470, 2003.

[6] C. Giraud-Carrier, "Metalearning-a tutorial," in *Tutorial at the 7th international conference on machine learning and applications (ICMLA), San Diego, California, USA*, 2008.

[7] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, *Metalearning: applications to data mining*. Springer Science & Business Media, 2008.

[8] E. Rendón, I. Abundez, A. Arizmendi, and E. Quiroz, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.

[9] I. Färber, S. Günnemann, H.-P. Kriegel, P. Kröger, E. Müller, E. Schubert, T. Seidl, and A. Zimek, "On using class-labels in evaluation of clusterings," in *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD*, 2010, p. 1.

[10] B. E. Dom, "An information-theoretic external cluster-validity measure," in *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2002, pp. 137–145.

[11] F. Kovács, C. Legány, and A. Babos, "Cluster validity measurement techniques," in *6th International symposium of hungarian researchers on computational intelligence*. Citeseer, 2005.

[12] Z. Ansari, M. Azeem, W. Ahmed, and A. V. Babu, "Quantitative evaluation of performance and validity indices for clustering the web navigational sessions," *World of Computer Science and Information Technology Journal*, vol. 1, no. 5, pp. 217–226, 2011.

[13] M. Kim and R. Ramakrishna, "New indices for cluster validity assessment," *Pattern Recognition Letters*, vol. 26, no. 15, pp. 2353–2363, 2005.

[14] C.-H. Chou, M.-C. Su, and E. Lai, "A new cluster validity measure and its application to image compression," *Pattern Analysis and Applications*, vol. 7, no. 2, pp. 205–220, 2004.

[15] H. Xiong, J. Wu, and J. Chen, "K-means clustering versus validation measures: a data-distribution perspective," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 39, no. 2, pp. 318–331, 2009.

[16] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "A study of some fuzzy cluster validity indices, genetic clustering and application to pixel classification," *Fuzzy Sets and Systems*, vol. 155, no. 2, pp. 191–214, 2005.

[17] B. S. S. M. zu Eissen and F. Wißbrock, "On cluster validity and the information need of users," *ACTA Press*, pp. 216–221, 2003.

[18] R. J. Hathaway and J. C. Bezdek, "Visual cluster validity for prototype generator clustering models," *Pattern Recognition Letters*, vol. 24, no. 9, pp. 1563–1569, 2003.

[19] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010, pp. 911–916.

[20] M. Meilă, "Comparing clusterings by the variation of information," in *Learning theory and kernel machines*. Springer, 2003, pp. 173–187.

[21] ——, "Comparing clusteringsan information based distance," *Journal of multivariate analysis*, vol. 98, no. 5, pp. 873–895, 2007.

[22] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[23] J. C. Bezdek and N. R. Pal, "Some new indexes of cluster validity," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 28, no. 3, pp. 301–315, 1998.

[24] B. Karrer, E. Levina, and M. E. Newman, "Robustness of community structure in networks," *Physical Review E*, vol. 77, no. 4, p. 046119, 2008.

[25] R. C. Tryon and D. E. Bailey, *Cluster analysis*. McGraw-Hill, 1970.

[26] J. M. Lewis, M. Ackerman, and V. De Sa, "Human cluster evaluation and formal quality measures: A comparative study," in *Proc. 34th Conf. of the Cognitive Science Society (CogSci)*, 2012, pp. 1870–1875.

[27] M. Hund, I. Färber, M. Behrisch, A. Tatu, T. Schreck, D. A. Keim, and T. Seidl, "Visual quality assessment of subspace clusterings," in *Proceedings of the ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics (IDEA)*, 2016, pp. 53–62.

[28] H. Rademacher, "On the partition function p(n)," *Proceedings of the London Mathematical Society*, vol. 2, no. 1, pp. 241–254, 1938.

[29] F. J. Brandenburg, A. Gleißner, and A. Hofmeier, "Comparing and aggregating partial orders with kendall tau distances," *Discrete Mathematics, Algorithms and Applications*, vol. 5, no. 02, p. 1360003, 2013.

[30] J. L. García-Lapresta and D. Pérez-Román, "Measuring consensus in weak orders," in *Consensual processes*. Springer, 2011, pp. 213–234.

[31] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. M. Pérez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243–256, 2013.

[32] G. W. Milligan, "An examination of the effect of six types of error perturbation on fifteen clustering algorithms," *Psychometrika*, vol. 45, no. 3, pp. 325–342, 1980.

[33] "Weka 3: Data mining software in java," http://http://www.cs.waikato.ac.nz/ml/weka/, accessed: 2016-01-26.

[34] S. Vassilvitskii and D. Arthur, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 2006, pp. 1027–1035.

[35] D. Pelleg, A. W. Moore *et al.*, "X-means: Extending k-means with efficient estimation of the number of clusters." in *ICML*, 2000, pp. 727–734.

[36] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[37] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[38] D. J. Rosenkrantz, R. E. Stearns, and P. M. Lewis, II, "An analysis of several heuristics for the traveling salesman problem," *SIAM journal on computing*, vol. 6, no. 3, pp. 563–581, 1977.

[39] L. Rokach and O. Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*. Springer, 2005, pp. 321–352.

[40] P. B. Brazdil, C. Soares, and J. P. Da Costa, "Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results," *Machine Learning*, vol. 50, no. 3, pp. 251–277, 2003.

[41] K. A. Smith-Miles, "Cross-disciplinary perspectives on meta-learning for algorithm selection," *ACM Computing Surveys (CSUR)*, vol. 41, no. 1, p. 6, 2008.

[42] Q. Sun, "Meta-learning and the full model selection problem," *PhD thesis, University of Waikato*, 2014.

[43] C. Castiello and A. M. Fanelli, "Computational intelligence for meta-learning: A promising avenue of research," in *Meta-Learning in Computational Intelligence*. Springer, 2011, pp. 157–177.

[44] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical association*, vol. 66, no. 336, pp. 846–850, 1971.

[45] M. Ackerman and S. Ben-David, "Clusterability: A theoretical study." in *AISTATS*, vol. 5, 2009, pp. 1–8.