

Using Internal Validity Measures to Compare Clustering Algorithms

Toon Van Craenendonck
Hendrik Blockeel

KU Leuven

Department of Computer Science

TOON.VANCRAENENDONCK@CS.KULEUVEN.BE

HENDRIK.BLOCKEEL@CS.KULEUVEN.BE

Abstract

Recently, significant effort has been made to automate the machine learning process in the context of supervised learning. This automation includes, amongst other things, the selection of an appropriate learning algorithm and corresponding hyperparameters for a particular learning problem. In contrast, such problems are much less studied for unsupervised tasks such as clustering. Nevertheless, users who want to cluster a data set are confronted with similar problems: a clustering algorithm should be selected from the wide variety of available algorithms, and usually some hyperparameters have to be set. In a supervised setting, model search is guided by performance measures that rely on known class labels, such as accuracy. However, these measures are not applicable to clustering as labels are usually not available. Instead, one might use internal validity measures that only rely on properties intrinsic to the data set. Several such measures are defined, and in this paper we study the usefulness of four of them for model selection. We perform experiments with these measures in combination with six clustering algorithms. While some measures are suited to use in hyperparameter optimization for some specific algorithms, we conclude that none of them is suited to compare across very different clustering algorithms.

1. Introduction

[Jain \(2010\)](#) defines clustering as the task of organizing data into sensible groups. Many other definitions can be found in the literature. Likewise, many different clustering algorithms exist, which may all produce very different partitions of the same data set. Even a single clustering algorithm can yield wildly different results depending on the parameter settings. A user who wants to cluster a data set is left with the difficult task of selecting an appropriate clustering algorithm and corresponding hyperparameters. Significant effort has been made to automate similar tasks in the context of supervised learning. For example, [Bergstra et al. \(2011\)](#) study hyperparameter optimization and [Brazdil et al. \(2003\)](#) study algorithm selection. The combined problem of doing both simultaneously was considered recently by [Thornton et al. \(2013\)](#). These approaches critically rely on the definition of a performance measure, capturing what it means for a model to be “good”. In supervised learning, measures such as accuracy and F-score are commonly accepted for this purpose. In clustering, however, we cannot use these measures as no class labels are available. Instead, internal validity measures can be used to assess the quality of a certain clustering of a given data set. Such measures capture ideas on properties of “good” clusterings, and can be calculated from only the data and the clustering under consideration. Several mea-

asures have been defined, and a first extensive experimental comparison was performed by Milligan and Cooper (1985). More recently, Arbelaitz et al. (2013) and Vendramin et al. (2010) performed similar experiments with an improved methodology and an updated set of validity measures. Their main goal was to evaluate the performance of these measures, by comparing the internal measures to external ones.

In this paper, we investigate the behaviour of four internal validity measures on clusterings generated with six very different clustering algorithms. We want to verify whether it would be useful to use any of the four tested internal measures as a performance measure in algorithm selection. If this would be the case, many of the existing techniques that are currently applied to supervised learning problems could be transferred to the unsupervised setting.

2. Validity measures

Until now, we only mentioned *internal* validity measures, as these are the ones that can be directly used in algorithm and parameter selection. Such measures rely only on properties intrinsic to the data to quantify the quality of a clustering. Examples include the silhouette, Davies-Bouldin and Caliński-Harabasz measures. A second category consists of the *external* measures, which compare a clustering to a given partition. Examples include the Rand and Jaccard measures.

2.1. External measures

It is important to note that in a typical clustering setting, we cannot rely on external measures to guide us in choosing an appropriate algorithm or good parameter settings, as we do not have a partition to compare to. However, external measures are often used to evaluate both clustering algorithms and internal validity measures. A common strategy is to cluster classification data sets, ignoring the class labels, and compare the produced partition to the known one using an external index. As discussed by Färber et al. (2010), this is often a flawed strategy. Nevertheless, keeping such limitations in mind, comparing to an external index can be useful. In our experiments we will use the Adjusted Rand Index (ARI), a modified version of the Rand Index (Rand, 1971), which can be thought of as the counterpart of accuracy for clustering.

2.2. Internal measures

Internal validity measures only rely on properties intrinsic to the data set. Most measures are based on the concepts of compactness (points in the same cluster should be similar) and separation (points in different clusters should be dissimilar). They differ in the way these concepts are defined, and how they are combined. We briefly discuss how these concepts are defined for each of the four internal measures that are used in the experiments, for a formal definition we refer to the respective papers. The within-cluster sum of squares, which is minimized by e.g. k-means, cannot be used as an internal validity measure because its value will decrease as the number of clusters increases, and reach the optimal value of zero for a solution in which every point is assigned to each own cluster.

- The **silhouette** index (Rousseeuw, 1987) defines compactness based on the pairwise distances between all elements in the cluster, and separation based on pairwise distances between all points in the cluster and all points in the closest other cluster.
- The **Davies-Bouldin** measure (Davies and Bouldin, 1979) defines compactness based on the distance of points in the cluster to its centroid, and separation based on distances between centroids.
- The **Caliński-Harabasz** measure (Caliński and Harabasz, 1974) also defines compactness based on the distance of points in a cluster to its centroid, and separation as the distance of the cluster centroid to the data centroid.
- The **Density-Based Cluster Validation** measure (Moulavi et al., 2014) transforms points to a space of reachability distances, hereby aiming to capture density properties of the data. Next, minimum spanning trees (MSTs) are constructed for each of the clusters, aiming to capture possibly non-convex cluster shapes. Cluster compactness is then defined as the maximal weight of the internal edges of the cluster MST, and separation between clusters is defined as the minimum reachability distance between the internal nodes of the cluster MSTs.

The first three measures are usually used in combination with the Euclidean distance, leading to a strong preference for spherical clusterings. The same holds for the large majority of existing internal validity measures. In contrast, the DBCV measure is based on MSTs and is able to also deal with non-convex cluster shapes. Moulavi et al. (2014) use DBCV to perform hyperparameter optimization for several density-based clustering algorithms.

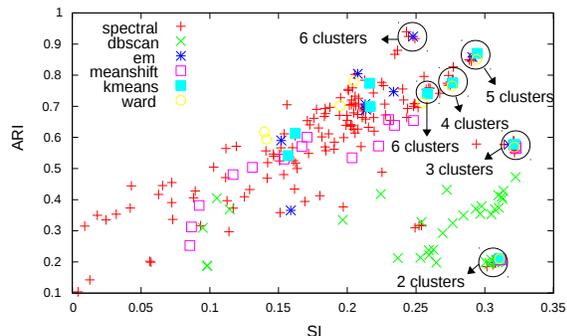
3. Clustering algorithms

Table 1 shows an overview of the algorithms used in the experiments, and the varied parameters. The parameter ranges were chosen to be wide enough to make sure that they contain values leading to a good solution. The algorithms were chosen as they are common representatives of very diverse types of clustering methods. For every data set and algorithm combination we generate solutions using a grid search, trying a maximum of 100 parameter combinations. For k-means, Ward and EM all numbers of clusters in the range were tried. For DBSCAN, spectral and meanshift the real-valued parameters were taken to be evenly spaced over the given intervals. Note that if we would just be interested in using one clustering algorithm, different strategies to select the algorithm parameters might be more appropriate. For example, Zelnik-Manor and Perona (2004) discuss automated ways to set the parameters of spectral clustering, including the number of clusters.

The actual complexity of running an algorithm for a particular data set in this setting becomes $k(C(A) + C(E))$ with k the number of tried parameter combinations, $C(A)$ the complexity of one run of algorithm A and $C(E)$ the complexity of evaluating the quality of one partition using evaluation measure E . For example, if we use k-means in combination with the silhouette index (which is $\mathcal{O}(N^2)$), the resulting time complexity is $\mathcal{O}(k(N + N^2)) = \mathcal{O}(N^2)$. This means that we spend much more time evaluating the clustering than producing it.

Algorithm	Complexity	Parameters
k-means	$\mathcal{O}(NK)$	K : # clusters
DBSCAN	$\mathcal{O}(N \log(N))$	ϵ : max. dist. to be nbs. $minPts$: # nbs. to be core pt
spectral	$\mathcal{O}(N^3)$	K : # clusters and k : # nbs. or σ : RBF scaling
Ward	$\mathcal{O}(N^2)$	K : # clusters
meanshift	$\mathcal{O}(N^2)$	RBF bandwidth
EM	$\mathcal{O}(NK)$	K : # clusters

Table 1: Algorithms used

Figure 1: SI vs. ARI for the *dermatology* data set

4. Results

In this section we investigate the relative abilities of the discussed clustering algorithms to score well on the validity measures. An important issue to consider when making such a comparison is the fact that DBSCAN and meanshift are able to identify points as noise. These points can actually be “true” noise, but don’t have to be: they are simply the points that agree with the algorithms’ definition of noise, under a certain parameter configuration. Following one of the strategies suggested by Moulavi et al. (2014), we remove noise points before calculating the SI, DB and CH measures and apply a proportional penalty. Such a strategy is not needed for DBCV. We have experimented with 27 UCI data sets, and in the remainder of this section we discuss some observations that were made during these experiments.

Assessing algorithms and internal measures using external measures can be misleading. While this is not a new observation (Färber et al., 2010), comparing with external measures is still a common strategy (Arbelaitz et al., 2013). One of the reasons why this can be misleading is illustrated in Figure 1. It shows that SI prefers 3-cluster solutions over the “true” 6-cluster solutions, that attain a much higher ARI. In these 3-cluster solutions, some clusters from the 6-cluster solution are merged. This suggests that the classes form a hierarchy, and that several cuts of the dendrogram are sensible to arrive at a good partitioning clustering. This is only one example of a situation in which comparing to an external measure can be misleading, Färber et al. (2010) provide an extensive discussion of this issue. Note that this does not mean that comparing to external labels cannot be useful, as they might still reflect how well one particular known clustering was reconstructed, but such comparisons should be made with these limitations in mind.

Highly imbalanced clusterings score well. Figure 2 illustrates that for the *sonar* data set, DBSCAN and meanshift are able to obtain significantly higher SI scores than the other algorithms. Similar behaviour was observed for several other data sets. Closer inspection shows that these high scoring clusterings are all very imbalanced, separating only a few points from all the others. This preference for imbalanced clusterings was also observed for other measures, e.g. illustrated for DBCV in Figure 4. This seems to be unwanted

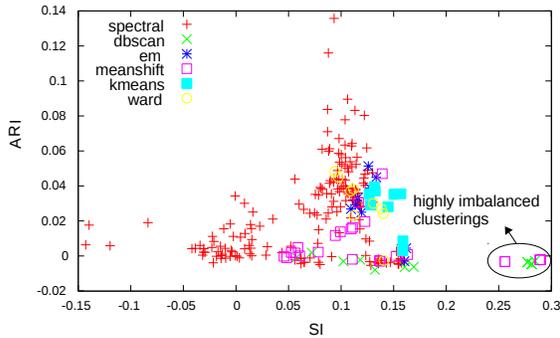


Figure 2: SI vs. ARI for the *sonar* data set

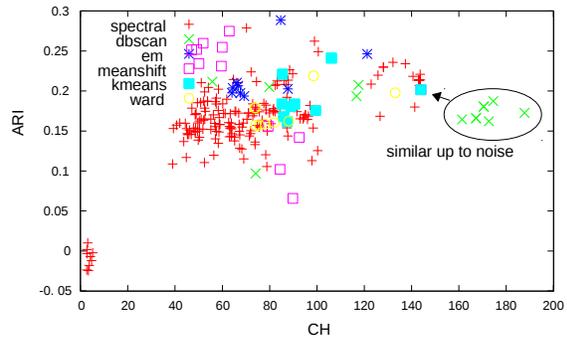


Figure 3: CH vs. ARI for the *glass* data set

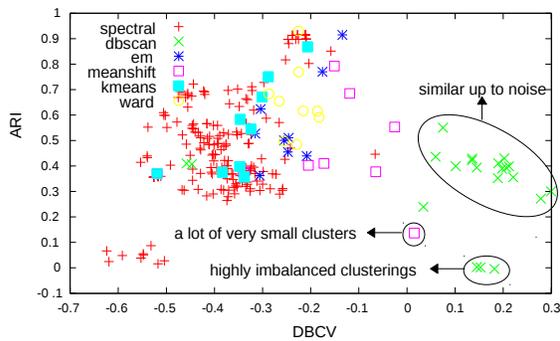


Figure 4: DBCV vs. ARI for the *wine* data

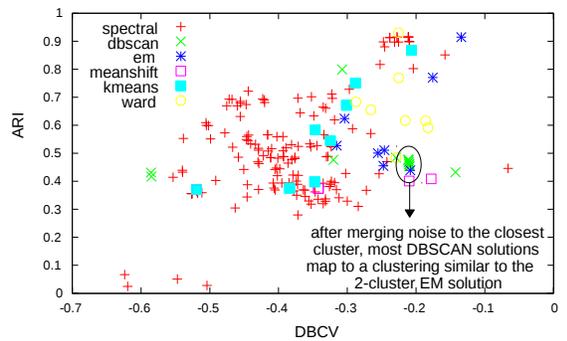


Figure 5: Figure 4 after merging noise and removing imbalanced clusterings

behaviour, as we are looking for interesting structure in the data set, and simply separating one or a few points from the others usually does not qualify as such.

All measures are heavily influenced by points identified as noise. Often the increase in the validity score due to identifying some points as noise is larger than the reduction of the score by applying a proportional penalty afterwards. Figures 3 and 4 illustrate this for the CH and DBCV measures. This allows algorithms able to identify noise to obtain much better scores, without actually finding more interesting structure. For the DBCV measure this effect is very severe, as meanshift and DBSCAN outperform all other algorithms on nearly all 27 data sets that we have considered. While this could be expected because of their similar assumptions about cluster structure, it is actually caused by the above mentioned effect of identifying noise points. We also experimented with the strategy of assigning each noise point to its closest cluster before calculating the DBCV score. Figure 5 shows the effect of this strategy for the *wine* data set. It illustrates that most clusterings that are generated by DBSCAN and that attain a relatively high DBCV score, are actually variations of the 2-cluster EM solution with a much lower score. Consequently, a significantly higher DBCV score does not necessarily indicate that the clustering identifies a very different structure.

We can simply use k-means to score well on the silhouette and Caliński-Harabasz measures. If we remove highly imbalanced clusterings (defined as the ones with $\frac{|c_{k-1}|}{|c_k|} < 0.1$, with c_k and c_{k-1} the largest and second-to-largest clusters, respectively) and solutions in which more than 50% of the points are identified as noise, most algorithms attain very similar maximal scores for the SI measure. In particular, k-means and spectral clustering score relatively well (as compared to the other algorithms) on almost all data sets. Overall, based on the results on the considered data sets, it seems reasonable to simply use k-means to produce clusterings with a good SI score. The same goes for the CH measure. Spectral clustering could also be used as it obtains very similar scores, but at a much higher computational cost. A similar conclusion could not be made for the DB and DBCV measures, as for these the sensitivity to noise and preference for imbalanced clusterings seem much more severe, rendering the “manual” filtering of imbalanced and noisy clusters more difficult. Consequently, it can be hard to determine whether a clustering attains a high score on these measures because it identifies interesting structure, or because it exploits these undesired properties.

5. Existing work on algorithm selection for clustering

De Souto et al. (2008); Soares et al. (2009); Ferrari and de Castro (2012) consider algorithm selection and ranking for clustering. They all rely on external measures to evaluate algorithms and construct rankings. However, typically we do not have external labels, and even if we do, using them is questionable (Färber et al., 2010). Recent work by Ferrari and Castro (2015) recognizes this shortcoming, and only relies on internal validity criteria to assess cluster quality (amongst which SI, CH and DB). However, our experiments suggest that most of these measures are not suited to compare between very different clustering algorithms. In particular, they suggest that one can simply use k-means to score well on the SI and CH measures.

6. Conclusion

A user who wants to cluster a data set is confronted with the difficult task of selecting an appropriate clustering algorithm and corresponding hyperparameters. Ideally, this search would be guided by a performance measure that allows to compare solutions generated by very different algorithms. In this paper, we have studied the applicability of four internal validity measures for this purpose. We have performed experiments with six clustering algorithms, aiming to provide insights into the validity measures and the ability of the algorithms to score well on these measures. We conclude that none of the four measures under consideration can be used to make a fair comparison between the six algorithms. All measures exhibit some undesired properties, of which users should be aware: sensitivity to points identified as noise, a preference for highly imbalanced solutions, or a bias towards spherical clusterings. This lack of an appropriate performance measure is the main obstacle to applying techniques for meta-learning and model selection from supervised learning to clustering.

Acknowledgements

Toon Van Craenendonck is supported by the Agency for Innovation by Science and Technology in Flanders (IWT).

References

- Olatz Arbelaitz, Ibai Gurrutxaga, Javier Muguerza, Jesús M. Pérez, and Iñigo Perona. An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1): 243–256, 2013.
- James S. Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyperparameter optimization. In *Advances in Neural Information Processing Systems 24*, pages 2546–2554. Curran Associates, Inc., 2011.
- Pavel B. Brazdil, Carlos Soares, and Joaquim Pinto da Costa. Ranking learning algorithms: Using ibl and meta-learning on accuracy and time results. *Machine Learning*, 50(3):251–277, 2003.
- Tadeusz Caliński and Joachim Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.
- David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227, 1979.
- Marcilio C. P. De Souto, Ricardo B. C. Prudêncio, Rodrigo G. F. Soares, Daniel S. A. De Araujo, Ivan G. Costa, Teresa B. Ludermir, and Alexander Schliep. Ranking and selecting clustering algorithms using a meta-learning approach. *Proceedings of the International Joint Conference on Neural Networks*, pages 3729–3735, 2008.
- Ines Färber, Stephan Günnemann, Hans-Peter Kriegel, Peer Kröger, Emmanuel Müller, Erich Schubert, Thomas Seidl, and Arthur Zimek. On Using Class-Labels in Evaluation of Clusterings. In *Proc. MultiClust Workshop in conjunction with 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2010), Washington, DC, USA*, 2010.
- Daniel G. Ferrari and Leandro Nunes De Castro. Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods. *Information Sciences*, 2015.
- Daniel G. Ferrari and Leandro Nunes de Castro. Clustering algorithm recommendation: A meta-learning approach. In *Swarm, Evolutionary, and Memetic Computing*, volume 7677 of *Lecture Notes in Computer Science*, pages 143–150. 2012.
- Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31: 651–666, 2010.
- Glenn W. Milligan and Martha C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179, 1985.

- Davoud Moulavi, Pablo A. Jaskowiak, R.J.G.B. Campello, Arthur Zimek, and Jörg Sander. Density-based clustering validation. In *Proceedings of the 14th SIAM International Conference on Data Mining (SDM), Philadelphia, PA*, 2014.
- William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.
- Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.
- RodrigoG.F. Soares, TeresaB. Ludermir, and FranciscoA.T. De Carvalho. An analysis of meta-learning techniques for ranking clustering algorithms applied to artificial data. In *Artificial Neural Networks ICANN 2009*, volume 5768 of *Lecture Notes in Computer Science*, pages 131–140. 2009. doi: 10.1007/978-3-642-04274-4_14.
- Chris Thornton, Frank Hutter, Holger H. Hoos, and Kevin Leyton-Brown. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, page 847–855, 2013.
- Lucas Vendramin, Ricardo J. G. B. Campello, and Eduardo R. Hruschka. Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, 3(4): 209–235, 2010.
- Lih Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608, 2004.