preprint - to appear at:
28th International Conference on Data Engineering (ICDE), Washington, DC, 2012

Evaluation of Clusterings – Metrics and Visual Support

Elke Achtert, Sascha Goldhofer, Hans-Peter Kriegel, Erich Schubert, Arthur Zimek

Institute for Informatics, Ludwig-Maximilians-Universität München, Germany
http://www.dbs.ifi.lmu.de
{achtert,goldhofer,kriegel,schube,zimek}@dbs.ifi.lmu.de

***Abstract*—When comparing clustering results, any evaluation metric breaks down the available information to a single number. However, a lot of evaluation metrics are around, that are not always concordant nor easily interpretable in judging the agreement of a pair of clusterings. Here, we provide a tool to visually support the assessment of clustering results in comparing multiple clusterings. Along the way, the suitability of a couple of clustering comparison measures can be judged in different scenarios.**

I. Introduction

Because of the potentially high complexity of data sets and in sight of different clustering results based on different parameters or different methods, as well as due to the low insight provided by numerical metrics, an abstract visual representation to give more information may be useful. Missing a *gold standard* for comparing clusterings (i.e., no perfect partitioning is known), methods for evaluating the results are required to find the best choice. Measure indices rate the similarity between clusterings, however with the large number of metrics known in the literature it is uncertain which measure index presents the best guess. Furthermore, we would be interested in reasons for the different clustering results and why one cluster algorithm performs better or worse on a given data set. To support research in these issues, we provide a tool evaluating measure indices and comparing clustering results. In order to apply a visualization for any clustering and any complexity of data, an abstract visualization needs to avoid visual cluttering.

Basically three categories in comparing clusterings exist [1]. The first is based on entropy or some other information measure, judging the mutual information between two clusterings (e.g. [2]). The second is based on set-matching, i.e., mapping each cluster from the first clustering to the most similar cluster in the second clustering and computing recall, precision or some other measure (e.g. [3]). The third is based on pair counting (e.g. [4]). Since pairing objects has a good potential to be visualized, our tool facilitates these possibilities. Our visual representation is therefore based on pair counting, resulting in better insight on the differences of clusterings while complementing common clustering evaluation metrics.

II. System

Trying to compare clustering results on the basis of data points and their assigned cluster often fails because we can not decide whether the assignment in one clustering is right or wrong compared to the other clustering result. Often, different assignments may be equally meaningful and could actually reveal hidden truths about a data set. How to evaluate clusterings in a meaningful manner remains an open research question [5], [6]. An approach to evaluation and to study common and different decisions between two (or more) clusterings is the examination of relationships of each pair of data points in the data set, resulting in a set of $\frac{n*(n-1)}{2}$ pairs for n data objects. Offering a description of each cluster by the pairs being clustered together results in a homogeneous set representing the similarity of those objects and comparing the assessment of similarity by different clusterings (or a clustering and a given ground truth). We visualize all these pairs as segments in a circle, as depicted schematically in Figure 1a. By assigning a different shade to each cluster, we paint each pair being represented by one segment of the circle according to its assigned cluster as in Figure 1c. Unshaded pairs (white) illustrate pairs not available in the clustering (i.e., that are separated in this clustering). This way, we visualize the size and the amount of clusters generated.

When using the pair representation of clusterings, comparing clusterings is based on comparing their common and their different sets of pairs. Therefore our visualization is extended by another ring in the circle visualization maintaining the order

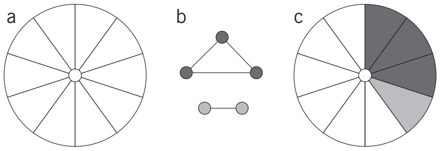

Fig. 1. a) all possible pairs drawn on a circle by dividing it into segments. b) set of five data points divided into two clusters and their representing pairs. c) pairs shaded to their corresponding cluster color.

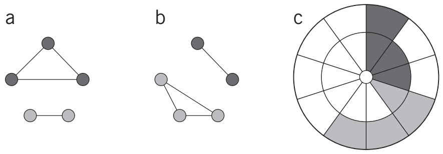

Fig. 2. Comparing cluster results of five data points. a) clustering 1. b) clustering 2. c) combination of two clusterings using our visualization where the inner circle represents clustering 1 (a), the outer circle clustering 2 (b).

TABLE I
CATEGORIES IN PAIRING CLUSTERED OBJECTS.

Clustering	Pairs in P	Pairs not in P
Pairs in Q	$a :=$ Pairs in both	$b :=$ Pairs in Q
Pairs not in Q	$c :=$ Pairs in P	$d :=$ Pairs in none

TABLE II
SOME WELL KNOWN PAIR COUNTING FORMULAS.

precision	*recall*	*F-measure*	*Rand*	*Jaccard*
$\frac{a}{a+c}$	$\frac{a}{a+b}$	$\frac{2a}{2a+b+c}$	$\frac{a+d}{a+b+c+d}$	$\frac{a}{a+b+c}$

of the pairs. Analogously to the first ring, the added clustering pairs are shaded according to their cluster color (as in Figure 2c). Each divided cluster results in white spaces representing missing pairs. Hence by the *circle segments visualization*, the fragmentation and the similarity of clusterings is visually accessible in an intuitive way.

Omitting the cluster association of the pairs, we have four categories: pairs that are common in both clusterings (completely shaded segments), pairs that do not exist in either clustering (completely white segments), pairs only present in the first clustering (segments with white space in second ring), and pairs present only in the second clustering (segments with white space in first ring). See Figure 2c and Table I. These categories are the working set of all pair counting measure indices. By counting the pairs in each category, we get an indicator for agreement and disagreement of the two clusterings. The well known asymmetric measures *precision* and *recall* are calculated as shown in Table II. For comparison of two clusterings without a given gold standard, symmetric measures like the *F-measure* (combining *precision* and *recall*) or the *Jaccard*-index [7] are more recommendable. The *Rand*-index [8], being defined as the ratio of the *pairs in both* and *the pairs in no clustering* to *all pairs*, states the probability that two objects are treated alike in both clusterings. In their evaluation of similarity measures, Pfitzner et al. [4] list over 40 different pair counting metrics, all giving different estimations for the similarity of two clusterings, requiring an evaluation of the metrics themselves.

Besides the variety of measure indices, the treatment of noise data points or noise clusters remains a field for research. Comparison of algorithms that have a notion of noise to algorithms that do not, requires adjustments in the evaluation process of pair counting to make a fair judgement between them. Our visualization displays the information in the process of creating the pairs being used by the measure indices before their data gets aggregated in the counting process. Therefore it provides us one picture as a reference for all pair counting metrics and the chance to investigate the fragmentation in detail, supporting comprehension and evaluation of the different clustering similarity values. Besides evaluating metrics, the proposed visualization supports the analysis and comparison of the clusterings itself.

The visualization (Figure 3) sorts pairs by the first cluster-result (inner ring), representing the reference clustering, then by the second result and so forth to group them into segments that are clustered the same way. The size of each segment is relative to its pair count, the red lines are inserted to emphasize the borders between primary segments. The complexity depends on the clustering results, but not on the actual data set size. This representation enables us to visually estimate the metrics by comparing the sizes of the involved segments. The complexity of the fragmentation and the dissimilarity of the clusterings is displayed. Adding hover interaction to the visualization highlights the selected cluster as demonstrated in Figure 4, enhancing the recognition of the distributed cluster.

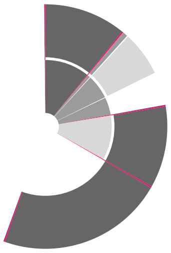

Fig. 3. Screenshot of the final *circle segments* visualization of two clusterings, each having three clusters.

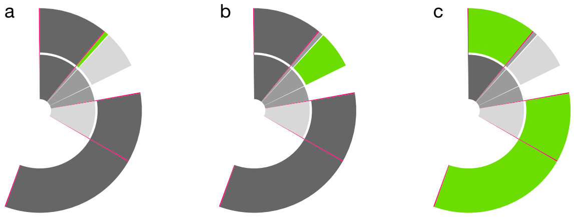

Fig. 4. Mouse hover on each segment and ring highlights the cluster.

Being an abstract visualization, clusters may be compared in the *circle segments visualization*. However, evaluating the actual quality of the clusterings requires a reference to the applied data set. An additionally complementing visualization representing the data set should be used in close conjunction. Therefore we have implemented the proposed visualization in the ELKI framework [9]–[12], along with release 0.5, that supplies a wide range of algorithms and visualizations. By adding, e.g., a scatterplot to our tool, as depicted in Figure 5, we are able to select segments of the visualization and highlight the corresponding data points in some visualization (here a simple scatterplot) of the data set. To track the selection, each segment is colored corresponding to the objects, allowing a deeper exploration of the cluster fragmentation in relation to the data set properties and human intuition.

Because we are mapping pairs to data points, when selecting a segment and highlighting it in the scatterplot the selection needs an adjustment. By selecting segments with pairs only existent in one clustering, one actually selects two different segments that describe the missing pairs in the other clustering. Thus, the selection seen in Figure 5 may be done by selecting each segment on its own or by selecting the last (largest) segment of the circle.

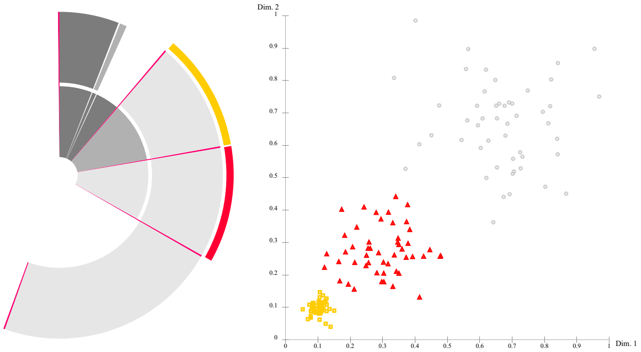

Fig. 5. Scatterplot with symbols representing the clusters of the reference clustering. The colors highlight pair segments the two clusterings agree upon. The reference clustering is EM ($k = 3; \delta = 0$), compared to (second ring) DBSCAN ($\varepsilon = 0.1; minPts = 5$).

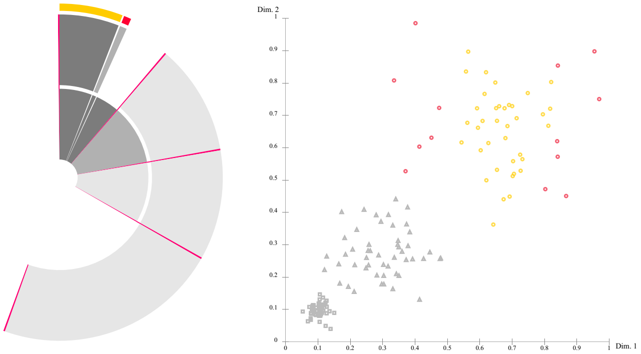

Fig. 6. Reference cluster being split by the compared clustering.

III. Demonstration Scenario

Here we sketch two examples[1] of clustering comparison. The first one compares EM-clustering [13] and DBSCAN [14] on a Gaussian data set with different densities and close clusters. The second example compares EM with k-means [15] on a data set with different densities and background scatter. Both data sets are very favorable to EM (which achieves near perfect results) and which we will therefore use in both scenarios as reference clustering on the inner ring.

In Figure 5, a comparison of EM and DBSCAN is shown. DBSCAN merged two clusters of EM (red and yellow) into one, producing a large pair segment nonexistent in the EM result (bottom segment). The corresponding scatterplot supports this analysis that for DBSCAN the two clusters were density-connected (and thus maybe the ε parameter was chosen too large). On the other hand, Figure 6 shows the first EM cluster selected, which is fragmented by DBSCAN into the denser core (yellow) of the cluster and a noise cluster (red) for the less dense outer objects (which could be included by choosing a larger ε). The missing pairs of DBSCAN are those connecting red and yellow objects. As stated in [4], a measure index should rate both clusterings as nearly equal since one clustering just merges two clusters of the other clustering. It could be worse, e.g., two clusters could be mixed up leading to a much stronger disagreement between both clusterings in terms of pairs. Comparing some related indices shown in Table III, we might judge the *Rand-index* or the *recall* as being the best guess for similarity, but the judgment solely based on the numerical measures remains inconclusive. Reasons for the cluster differences lie in DBSCAN having problems with different densities in the data set (see the discussion in [16]); it cannot be parametrized to detect all clusters correctly at the same time, while Gaussian clusters are expected to being easily modeled by the EM approach.

TABLE III
Resulting indices of EM compared to DBSCAN

precision	0.5616992582602832
recall	0.9066666666666666
F-measure	0.6936608722806287
Rand	0.7366442953020135
Jaccard	0.5309960159362550

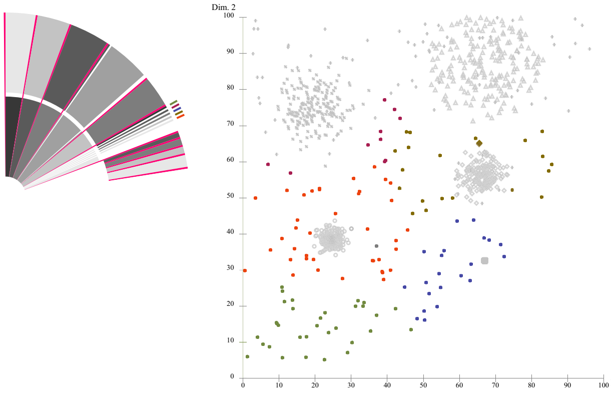

Fig. 7. Reference clustering is EM ($k = 6; \delta = 0$), second ring is k-means ($k = 6; \delta = 0$).

The second scenario uses a data set with different densities and background scatter. EM clustering does not have a special treatment of noise, however increasing k by 1 provides good results: EM indeed produces a cluster with low density and high variance that contains most noise objects. The same trick cannot work for k-means, so we chose the exact k there for best results. Because of the number of clusters and their size, about three quarters of pairs exist in neither clustering as seen in Figure 7. There is a strong agreement in most clusters, resulting in large shared segments and thus a high similarity value in many popular measure indices (Table IV). Inspecting the segment containing most of the background scatter, we can see that k-means does not really handle noise, which is distributed to the various actual clusters. By hovering over the larger segments in the second ring, the highlights show that they are part of one of the larger EM clusters except for one tiny segment containing the noise object pairs in the corresponding EM cluster and an unmatched segment containing the pairs with one object in the cluster and the other in the background scatter. This is a known defect of k-means clustering that it will assign such noise objects to the nearest cluster found. To increase readability, the segment containing

[1] Example data sets available at http://elki.dbs.ifi.lmu.de/

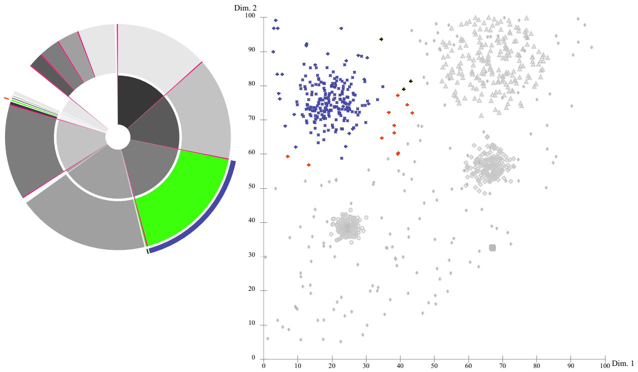

Fig. 8. Hiding of unclustered pairs increases readability in more complex scenarios.

TABLE IV
RESULTING INDICES OF EM COMPARING TO k-MEANS

precision	0.8185578744994065
recall	0.944720079641935
F-measure	0.881337059843367
Rand	0.9558548790658883
Jaccard	0.9545458224405721

nonexistent pairs can be hidden, as shown in Figure 8. Here the problem treating noise data points becomes more visible. The scatterplot shows the EM cluster in blue and the larger extent of the k-means cluster (red and olive objects).

This scenario demonstrates the ambiguity of handling noise by clustering algorithms that do not have an explicit notion of noise. EM adapts surprisingly well to this challenge when given an extra cluster to use. The same trick did not work for k-means, which just includes noise objects in the nearest cluster. The corresponding measure indices in Table IV however rate both clusterings as being more or less equal, but with a higher amount of noise the rating would likely be worse. Nonetheless, the reasons for their differences are identified by means of the visualizations.

IV. CONCLUSIONS

We have introduced our visualization tool and demonstrated the application on two clustering comparison examples. As demonstrated, our *circle segments visualization* supports clustering comparison as well as metrical evaluation in an accessible approach. The visualization is capable of a quantitative visualization of the differences in clustering results. Since its complexity depends on the dissimilarity and number of clusters only, it can also be used for large and high-dimensional data sets that cannot be easily visualized by other means and still provide the analyst with useful insights such as clusters being merged or split.

The proposed visualization can also be expanded to multiple clusterings as shown in Figure 9. The circular layout is beneficial here, since fragmentation increases in the outer circles. In general all possible pair subsets are visualized, resulting in a unified representation for all pair counting measures and still having a higher information value than the numerical measures themselves while also supporting further exploration of the details. The interactive visualization using segments enables the researcher to obtain knowledge about the relation of clusters to each other, such as clusters being merged, divided or distributed into multiple fragments by the other result. In particular large pair segments that are missing from one result but present in the other are indicative of such structural differences in the cluster results.

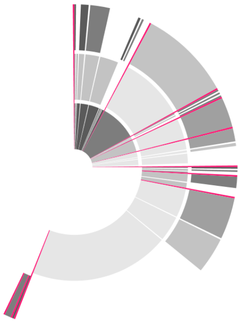

Fig. 9. Visualizing more than 2 clusterings.

This evaluation tool comes along with ELKI release 0.5, available at: http://elki.dbs.ifi.lmu.de.

REFERENCES

[1] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clustering comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, 2010.

[2] M. Meila, "Comparing clusterings – an axiomatic view," in *Proc. ICML*, 2005.

[3] S. Günnemann, I. Färber, E. Müller, I. Assent, and T. Seidl, "External evaluation measures for subspace clustering," in *Proc. CIKM*, 2011.

[4] D. Pfitzner, R. Leibbrandt, and D. Powers, "Characterization and evaluation of similarity measures for pairs of clusterings," *KAIS*, vol. 19, no. 3, pp. 361–394, 2009.

[5] 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 *Proc. ACM SIGKDD Workshop MultiClust*, 2010.

[6] H.-P. Kriegel, E. Schubert, and A. Zimek, "Evaluation of multiple clustering solutions," in *Proc. ECML PKDD Workshop MultiClust*, 2011.

[7] P. Jaccard, "Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines," *Bulletin de la Societe Vaudoise des Sciences Naturelles*, vol. 37, pp. 241–272, 1901.

[8] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *JASA*, vol. 66, no. 336, pp. 846–850, 1971.

[9] E. Achtert, H.-P. Kriegel, and A. Zimek, "ELKI: a software system for evaluation of subspace clustering algorithms," in *Proc. SSDBM*, 2008.

[10] E. Achtert, T. Bernecker, H.-P. Kriegel, E. Schubert, and A. Zimek, "ELKI in time: ELKI 0.2 for the performance evaluation of distance measures for time series," in *Proc. SSTD*, 2009.

[11] E. Achtert, H.-P. Kriegel, L. Reichert, E. Schubert, R. Wojdanowski, and A. Zimek, "Visual evaluation of outlier detection models," in *Proc. DASFAA*, 2010.

[12] E. Achtert, A. Hettab, H.-P. Kriegel, E. Schubert, and A. Zimek, "Spatial outlier detection: Data, algorithms, visualizations," in *Proc. SSTD*, 2011.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc. B*, vol. 39, no. 1, pp. 1–31, 1977.

[14] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. KDD*, 1996.

[15] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *5th Berkeley Symposium on Mathematics, Statistics, and Probabilistics*, vol. 1, 1967, pp. 281–297.

[16] H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *WIREs DMKD*, vol. 1, no. 3, pp. 231–240, 2011.