CICE-BCubed: A New Evaluation Measure for Overlapping Clustering Algorithms

Henry Rosales-Méndez
Computer Science
Departament
Universidad de Oriente, Cuba
henry@csd.uo.edu.cu

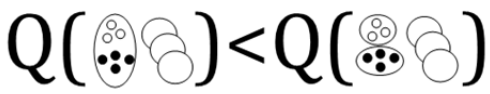
Yunior Ramírez-Cruz
Center for Pattern
Recognition and Data Mining
Universidad de Oriente, Cuba
yunior@cerpamid.co.cu

Abstract

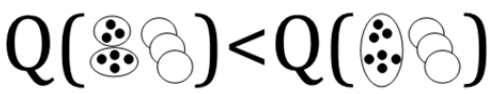
The evaluation of clustering algorithms is a field of Pattern Recognition still open to extensive debate. Most quality measures found in the literature have been conceived to evaluate non-overlapping clusterings, even when most real-life problems are better modeled using overlapping clustering algorithms. A number of desirable conditions to be satisfied by quality measures used to evaluate clustering algorithms have been proposed, but measures fulfilling all conditions still fail to adequately handle several phenomena arising in overlapping clustering. In this paper, we focus on a particular case of such desirable conditions, which existing measures that fulfill previously enunciated conditions fail to satisfy. We propose a new external evaluation measure that correctly handles the studied phenomenon for the case of overlapping clusterings, while still satisfying the previously existing conditions.

Previous Work

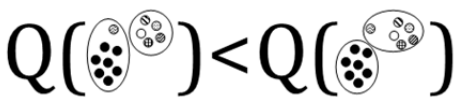
Several authors have enunciated sets of conditions aiming to assess the convenience of using specific evaluation measures. Amigó et al. [1], after conducting an extensive survey of existing conditions, summarized them into a set of four conditions and proved that all previous conditions were covered by these.



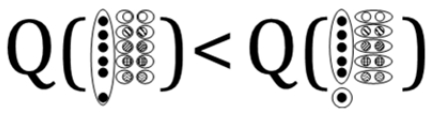
a) Homogeneity condition.



b) Completeness condition.



c) Rag bag condition.



d) Clusters size versus quantity condition.

Figure 1. Set of conditions proposed by Amigó et al.

Upon the presentation of their four conditions, Amigó et al. conducted an extensive study on a large number of existing evaluation measures to determine the extent to which they satisfy the proposed conditions. They concluded that the BCubed F_α measure is the sole evaluation measure that satisfies all four conditions. Since BCubed is defined for non-overlapping clustering, Amigó et al. propose Extended BCubed (3), an extension of BCubed suited for evaluating overlapping clusterings, which contains BCubed as an special case when zero overlapping is present.

$$P = \frac{1}{|U|}\sum_{o\in U}\frac{1}{|\bigcup_{g\in G(o)} g|}\sum_{o'\in E(o,G)}\frac{min(|G(o)\cap G(o')|,|C(o)\cap C(o')|)}{|G(o)\cap G(o')|} \quad (1)$$

$$R = \frac{1}{|U|}\sum_{o\in U}\frac{1}{|\bigcup_{g\in C(o)} g|}\sum_{o'\in E(o,C)}\frac{min(|G(o)\cap G(o')|,|C(o)\cap C(o')|)}{|C(o)\cap C(o')|} \quad (2)$$

$$F_\alpha(P,R) = \frac{1}{\alpha(\frac{1}{P})+(1-\alpha)(\frac{1}{R})} \quad (3)$$

Where U represents the collection, G stands for the candidate clustering, C for the gold standard, $G(o)$ represents the set of candidate clusters containing object o, $C(o)$ is the set of classes of the gold standard containing o, $E(o,G)$ is the set of objects co-occurring with o in at least one candidate cluster, and $E(o,C)$ is the set of objects co-occurring with o in at least one class of the gold standard. The sets of objects yielded by $E(o,G)$ and $E(o,C)$ contain object o itself.

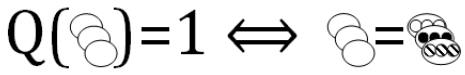
Unlike, the Extended BCubed measures do not rely on directly calculating the amount of set-matching between classes and candidate clusters. Instead, they analyze the set of object pairs and consider the decisions of placing pairs together or not, with respect to the gold standard. Here, we focus on a problem pointed out by Amigó et al., namely the fact that the maximum Extended BCubed F_α score may be obtained when evaluating a candidate clustering that is not identical to the gold standard, as shown in the following example:

Candidate	*Gold*
$C_1 : 1,2,4$	$G_1 : 1,3,4$
$C_2 : 1,3$	$G_2 : 1,2$
$C_3 : 4,3$	$G_3 : 4,2$
$C_4 : 2,5$	$G_4 : 3,5$
$C_5 : 3,5,6$	$G_5 : 2,5,6$
$C_6 : 2,6$	$G_6 : 3,6$

Our Proposal

We will treat this desired behavior as a supplementary condition, which we will refer to as the *Perfect match condition*, and is formally enunciated as follows:

Perfect match condition: an evaluation measure must yield the maximum score for a candidate clustering if and only if it is identical to the gold standard.



When evaluating non-overlapping clusterings, most of the existing evaluation measures satisfy the perfect match condition. However, when overlapping clusterings are involved, it is a challenge for a measure to fulfill that condition. We propose a new family of evaluation measures: *Cluster-Identity-Checking Extended BCubed* (*CICE-BCubed* for short). Analogous to the BCubed and the Extended BCubed families, CICE-BCubed consists in a new way to calculate precision, recall and the F-measure.

$$\psi_B(A_i) = B_j \in B \ such\ that\ \left[sim(A_i,B_j) = \max_k sim(A_i,B_k)\right] \quad (4)$$

$$Jaccard(A_i,B_j) = \frac{|A_i\cap B_j|}{|A_i\cup B_j|} \quad (5)$$

$$\Phi(o,o',A,B) = \frac{1}{|A(o,o')|}\sum_{A_i\in A(o,o')} sim(A_i,\psi_B(A_i)) \quad (6)$$

$$\hat{P} = \frac{1}{|U|}\sum_{o\in U}\frac{1}{|\bigcup_{g\in G(o)} g|}\sum_{o'\in E(o,G)}\frac{min(|G(o)\cap G(o')|,|C(o)\cap C(o')|)\cdot\Phi(o,o',G,C)}{|G(o)\cap G(o')|} \quad (7)$$

$$\hat{R} = \frac{1}{|U|}\sum_{o\in U}\frac{1}{|\bigcup_{g\in C(o)} g|}\sum_{o'\in E(o,C)}\frac{min(|G(o)\cap G(o')|,|C(o)\cap C(o')|)\cdot\Phi(o,o',C,G)}{|C(o)\cap C(o')|} \quad (8)$$

$$\hat{F}_\alpha(\hat{P},\hat{R}) = \frac{1}{\alpha(\frac{1}{\hat{P}})+(1-\alpha)(\frac{1}{\hat{R}})} \quad (9)$$

The measures of the CICE-BCubed family have a considerably high worst-case time complexity, $O(n^3 \log n)$, where n is the number of objects in the collection, for the case where the candidate clustering has n clusters, each containing $n-1$ objects.

Conclusions

We have proposed CICE-BCubed, a new family of evaluation measures for clustering algorithms, which correctly handle phenomena arising in the evaluation of overlapping clusterings that are inconveniently handled by previously existing measures.

References

[1] Amigó, E.; Gonzalo, J.; Artiles, J.; Verdejo, F. A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12(4), 461-486 (2009)