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

Henry Rosales-Méndez[1] and Yunior Ramírez-Cruz[2]

[1] Computer Science Department
Universidad de Oriente, Santiago de Cuba, Cuba
henry.rosales@csd.uo.edu.cu
[2] Center for Pattern Recognition and Data Mining
Universidad de Oriente, Santiago de Cuba, 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 evaluation measure that correctly handles the studied phenomenon for the case of overlapping clusterings, while still satisfying the previously existing conditions.

1 Introduction

Clustering is one of the most widely investigated problems in Pattern Recognition. It consists on separating an object collection into a set of clusters, generally attempting to place similar objects in a common cluster and dissimilar objects in distinct clusters. Partitional clustering algorithms split the collection into a set of disjoint clusters, in such a way that an object may belong to only one cluster. On the other hand, overlapping clustering algorithms allow objects to belong to multiple clusters.

Evaluating the quality of the outcomes of clustering algorithms is a necessary, yet challenging task, for which a large number of evaluation measures have been proposed. In the literature, evaluation measures are divided into external, relative or internal. Some authors refer to external measures as extrinsic and to internal measures as intrinsic, respectively. Internal measures asses the quality of a clustering by analyzing intra- and inter-cluster properties, without considering any external knowledge. Relative measures compare the results of multiple runs of a clustering algorithm, such runs differring in the parameter combinations used. On the other hand, external measures compare candidate clusterings to a

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part I, LNCS 8258, pp. 157–164, 2013.

gold standard, i.e. a handcrafted set of known clusters built by human experts taking into account the characteristics of the collection. In this work, we focus on external evaluation measures.

Due to the abundance of evaluation measures, some authors have devoted a considerable effort to enunciate desirable conditions that evaluation measures are expected to satisfy. While these conditions are inherently intuitive, and no universal consensus is likely to be obtained regarding their validity, extensive argumentation has been offered regarding their usefulness for providing criteria to consider when choosing an evaluation measure to assess the quality of clustering algorithms under different real-life and laboratory conditions.

Generally, most existing conditions have been enunciated for the particular case of partitional clustering, hinting to their extensibility to overlapping clustering. However, situations arrive when evaluation measures that have been proved to fulfill a wide range of conditions for partitional clusterings behave in an undesired manner when applied to overlapping clusterings. In this work, we take as starting point the work by Amigó et al. [1], who enunciate four conditions, prove them to cover all previously enunciated conditions, and show that BCubed F_α [2] is the sole to fulfill all four conditions. Amigó et al. propose Extended BCubed as an extension to BCubed for the case of overlapping clusterings. Here, we target a specific problem that is inadequately handled by Extended BCubed, namely that of assigning the optimum score to clusterings that are not identical to the gold standard. We formalize this desired behavior as a condition and propose CICE BCubed, a new extension to Extended BCubed, that satisfies the new condition while maintaining the established good characteristics of its predecessor.

The remainder of this paper is organized as follows. In Section 2 we briefly review existing work, focusing on the sets of conditions that have been previously enunciated, as well as the Extended BCubed family of evaluation measures. We describe our proposals in Section 3, and, finally, present our conclusions in Section 4.

2 Previous Work

Several authors have enunciated sets of conditions aiming to assess the convenience of using specific evaluation measures. Meila [3] enunciated twelve properties that were satisfied by the evaluation measure Variation Information. Later on, other authors used these properties as a set of conditions to be satisfied by other quality measures. Dom [4] proposed a distinct set of five conditions, which were later extended to seven by Rosenberg [5].

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.

All of the four conditions proposed by Amigó et al. are expressed as situations under which a clustering D_1, which is considered to be worse than a clustering D_2, is expected to be given a worse score. An evaluation measure is considered

to satisfy the condition if it behaves in this expected manner for all cases where such a situation arises. The conditions proposed by Amigó et al. are enunciated as follows:

- *Homogeneity*: Let D_1 be a clustering where one cluster G_k contains objects belonging to two classes[1]: C_i and C_j. Let D_2 be a clustering identical to D_1, except for the fact that instead of the cluster G_k, it contains two clusters G'_{k_1} and G'_{k_2}, one of them containing only objects belonging to C_i and the other containing only objects belonging to C_j. An evaluation measure satisfying the homogeneity condition should score D_1 worse than D_2.
- *Completeness*: Let D_1 be a clustering where two clusters G_1 and G_2 contain only objects belonging to one class C_k. Let D_2 be a clustering identical to D_1, except for the fact that instead of the clusters G_1 and G_2, it contains one cluster $G_{1,2}$, which is the union of G_1 and G_2. An evaluation measure satisfying the completeness condition should score D_1 worse than D_2.
- *Rag Bag*: Let D_1 be a clustering where one cluster G_{clean} contains n objects belonging to one class C_i plus one object belonging to a different class C_j and one cluster G_{noise} contains n objects belonging to n distinct classes. Let D_2 be a clustering identical to D_1, except for the fact that the object in G_{clean} that does not belong to the same class as all other objects is placed instead in G_{noise}. An evaluation measure satisfying the rag bag condition should score D_1 worse than D_2.
- *Clusters size versus quantity*: Let D be a clustering where one cluster G_{large} contains $n+1$ objects belonging to one class C_1 and n clusters G_1, G_2, ..., G_n, contain each on two objects belonging to the same class. Let D_1 be a clustering identical to D, except for the fact that instead of the two-object clusters G_1, G_2, ..., G_n, it contains $2n$ unary clusters containing the corresponding objects. Let D_2 be a clustering identical to D, except for the fact that instead of the cluster G_{large}, it contains one cluster of size n and one cluster of size 1. An evaluation measure satisfying the clusters size versus quantity condition should score D_1 worse than D_2.

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 [2] is the sole evaluation measure that satisfies all four conditions. Since BCubed is defined for non-overlapping clustering, Amigó et al. propose Extended BCubed, an extension of BCubed suited for evaluating overlapping clusterings, which contains BCubed as an special case when zero overlapping is present.

[1] Amigó et al. and other authors refer to the clusters of the gold standard clustering as *classes* or *categories*. Here, we will follow this terminology convention for simplicity.

The Extended BCubed family builds on the traditional Information Retrieval triad of evaluation measures Precision, Recall, F-measure [6]. Unlike these, 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. Extended BCubed precision evaluates the amount to which the decisions made by the evaluated algorithm of placing pairs of objects together in one or several clusters are correct, whereas Extended BCubed Recall evaluates the amount to which the evaluated algorithm is capable of putting together the pairs of objects that co-occur in classes of the gold standard. As in the case of the traditional IR measures, the Extended BCubed F-measure provides a trade-off between Extended BCubed precision and Extended BCubed recall.

The Extended BCubed precision is defined as

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

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.

In a similar manner, Extended BCubed recall is defined as

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

whereas the Extended BCubed F-measure is defined as

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

The authors propose to use $\alpha = 0.5$ so F_α behaves as the harmonic mean between precision and recall. The original BCubed measures analyze the fact that pairs of objects are placed together or not in clusters and/or classes. To adjust to the overlapping clustering case, the Extended BCubed measures additionally analyze the number of clusters and/or classes in which pairs of documents are placed together.

3 Our Proposal

Starting from the premise that the four conditions proposed by Amigó et al. are the most complete set of conditions, and the fact that the Extended BCubed

family is the overlapping clustering-oriented extension of the BCubed family, out of which BCubed F_α was proved to be the sole that satisfies all three conditions, we take Extended BCubed as the basis for further amelioration.

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

The reason why Extended BCubed F_α yields the maximum score for these cases is that it only checks for the number of clusters and/or classes where object pairs co-occur, but at no point attempt to establish a mapping between the set of candidate clusters and the set of classes. Such mapping would allow to determine whether the set of clusters where an object pair co-occurr in the candidate clustering is equivalent to the set of classes where they co-occurr in the gold standard.

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. Being Extended BCubed F_α the sole measure that satisfies the initial four conditions, we take it as a starting point to propose a new extension that, while maintaining the desirable characteristics of Extended BCubed, also satisfies the perfect match 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.

Being an extension of Extended BCubed, CICE-BCubed works by analyzing the object pairs that co-occur in clusters and/or classes. Unlike its predecessor, the measures of the CICE-BCubed family establish a mapping between the set of candidate clusters and the classes of the gold standard, in such a way that the candidate clustering's respect of that matching is evaluated along with the number of co-occurrences of object pairs. To do so, we introduce the *Cluster*

Identity Index (*CII* for short), a factor $\Phi(o_1, o_2, A, B)$ that yields values in the interval $[0, 1]$. For a pair of objects, the CII estimates the degree of similarity of all the clusters in A to their most similar class in B containing the pair. To define the CII, we use the auxiliary function $\psi_B(A_i)$ that determines the cluster $B_j \in B$ that best matches A_i, as follows:

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

where *sim* represents some function that calculates how similar two clusters are. Here, we calculate cluster similarity using Jaccard's index [7], which is defined as:

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

We chose Jaccard's index because it only yields the maximum score for two identical clusters. Other functions displaying the same behavior may as well be used, e.g. Rand's coefficient or the traditional IR F-measure.

When comparing a candidate clustering A to the gold standard B, for a pair of objects belonging to A, the CII averages the similarity values of clusters $A_i \in A$ that contain the pair to their best matching classes $B_j \in B$. Considering $A(o, o')$ as the set of all the clusters in A that contain the pair of objects (o, o'), the CII is defined as

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

The CII will yield the maximal value only if the best matching classes for every cluster in $A_i \in A$ are identical to their corresponding clusters. The aforementioned auxiliary functions are used for defining the measures of the CICE-BCubed family. CICE-BCubed precision is defined as

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

whereas CICE-BCubed recall is defined as

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

and the CICE-BCubed F-measure is defined as

$$\hat{F_\alpha}(\hat{P}, \hat{R}) = \frac{1}{\alpha(\frac{1}{\hat{P}}) + (1 - \alpha)(\frac{1}{\hat{R}})} \tag{9}$$

In order to keep the desirable characteristics of Extended BCubed F_α, both CICE-BCubed precision and recall maintain the terms from their Extended

BCubed homologous, but in both cases the terms in the numerators are multiplied by the factor CII, which prevents them from yielding optimal values for candidate clusterings that are not identical to the gold standard.

We will now analyze the behavior of the CICE-BCubed family of evaluation measures. Firstly, it is straightforward that CICE-BCubed precision and recall always obtain a maximum score for a candidate clustering which is identical to the gold standard. In this case, when calculating the CII, every cluster is always mapped to the class of the gold standard that is identical to it, thus always yielding the maximum value. Since the portions of Equations 7 and 8 inherited from Extended BCubed also contribute a maximum score, both CICE-BCubed precision and recall yield the maximum score and, consequently, so does the CICE-BCubed F-measure.

Secondly, we will demonstrate, using proof by contrapositive, that the fact of obtaining a maximal score of CICE-BCubed precision and recall implies that the candidate clustering is identical to the gold standard. Let A be a candidate clustering, which is not identical to the gold standard B. Under this condition, there must be at least one cluster $A_i \in A$ whose best matching class is not identical to it. For object pairs occurring in such cluster A_i, the CII will not yield the maximum score, thus preventing CICE-BCubed precision and CICE-BCubed recall, as defined in Equations 7 and 8, from yielding the maximum score. If CICE-BCubed precision and recall do not yield the maximum score, neither does the CICE-BCubed F-measure. Thus, we have proven that evaluating a candidate clustering which is not identical to the gold standard yields non-maximal CICE-BCubed precision, recall and F-measure, which, in turn, demonstrates that obtaining maximal CICE-BCubed precision, recall and F-measure implies that the evaluated candidate clustering is identical to the gold standard.

As a consequence of the previous proofs, we may conclude that CICE-BCubed F_α satisfies the perfect match condition. Additionally, it inherits from Extended BCubed F_α the behavior that satisfies the original four conditions enunciated by Amigó et al., which is not modified by the CII factor. This factor, while always causing the measure to yield values that are at most equal to the equivalent Extended BCubed F_α, does not alter the orientation of the inequalities that prove that Extended BCubed F_α satisfies the four conditions [1].

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. However, taking into account that evaluation is generally performed as an offline task during the process of tuning an algorithm for practical application, we consider that this time complexity is affordable given the benefits of relying on more robust evaluation measures.

4 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.

We took as a starting point the four conditions enunciated by Amigó et al., as well as the Extended BCubed F_α measure, which is reported to be the sole measure that satisfies the initial four conditions, and attacked one of the known problems that it faces when used for overlapping clusterings, namely that of assigning the maximum score to candidate clusterings that are not identical to the gold standard. We prove that our proposed counterpart, CICE-BCubed F_α, does handle this situation adequately, while continuing to satisfy the previous four conditions.

It should be noted nonetheless that the four conditions proposed by Amigó et al., as well as any other set of conditions, do not necessarily enjoy universal acceptation. Because of that, absolute statements regarding whether a particular evaluation measure should be considered better than others may not be appropriate. However, we consider that the existing conditions, as well as the new condition we treated in this paper, do reflect desirable characteristics of clustering evaluation measures, thus supporting the strength of the proposed measures and the convenience of their use.

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)
2. Bagga, A., Baldwin, B.: Entity-based cross-document coreferencing using the vector space model. In: Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and the 17th International Conference on Computational Linguistics, pp. 79–85 (1998)
3. Meilă, M.: Comparing Clusterings by the Variation of Information. In: Schölkopf, B., Warmuth, M.K. (eds.) COLT/Kernel 2003. LNCS (LNAI), vol. 2777, pp. 173–187. Springer, Heidelberg (2003)
4. Dom, B.: An information-theoretic external cluster-validity measure. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pp. 137–145 (2002)
5. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 410–420 (2007)
6. van Rijsbergen, C.: Foundation of evaluation. Journal of Documentation 30(4), 365–373 (1974)
7. Halkidi, M., Batistakis, Y., Vazirgiannis, M.: On Clustering Validation Techniques. Journal of Intelligent Information Systems 17(2-3), 107–145 (2001)