Pattern Recognition 65 (2017) 58–70

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Ground truth bias in external cluster validity indices

Yang Lei[a,*], James C. Bezdek[a], Simone Romano[a], Nguyen Xuan Vinh[a], Jeffrey Chan[b], James Bailey[a]

[a] Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia
[b] School of Science (Computer Science and Information Technology), RMIT University, Victoria, Australia

ARTICLE INFO

Keywords:
External cluster validity indices
Rand index
Ground truth bias
Quadratic entropy

ABSTRACT

External *cluster validity indices* (CVIs) are used to quantify the quality of a clustering by comparing the similarity between the clustering and a ground truth partition. However, some external CVIs show a biased behavior when selecting the most similar clustering. Users may consequently be misguided by such results. Recognizing and understanding the bias behavior of CVIs is therefore crucial.

It has been noticed that, some external CVIs exhibit a preferential bias towards a larger or smaller number of clusters which is monotonic (directly or inversely) in the number of clusters in candidate partitions. This type of bias is caused by the functional form of the CVI model. For example, the popular *Rand Index* (RI) exhibits a monotone increasing (NCinc) bias, while the *Jaccard Index* (JI) index suffers from a monotone decreasing (NCdec) bias. This type of bias has been previously recognized in the literature.

In this work, we identify a new type of bias arising from the distribution of the ground truth (reference) partition against which candidate partitions are compared. We call this new type of bias *ground truth* (GT) bias. This type of bias occurs if a change in the reference partition causes a change in the bias status (e.g., NCinc, NCdec) of a CVI. For example, NCinc bias in the RI can be changed to NCdec bias by skewing the distribution of clusters in the ground truth partition. It is important for users to be aware of this new type of biased behavior, since it may affect the interpretations of CVI results.

The objective of this article is to study the empirical and theoretical implications of GT bias. To the best of our knowledge, this is the first extensive study of such a property for external CVIs. Our computational experiments show that 5 of 26 pair-counting based CVIs studied in this paper, which are all functions of the RI, exhibit GT bias. Following the numerical examples, we provide a theoretical analysis of GT bias based on the relationship between the RI and quadratic entropy. Specifically, we prove that the quadratic entropy of the ground truth partition provides a computable test which predicts the NC bias status of the RI.

1. Introduction

Clustering is one of the fundamental techniques in data mining, which helps users explore potentially interesting patterns in unlabeled data. Cluster analysis has been widely used in many areas, ranging from bioinformatics [1] and market segmentation [2] to information retrieval [3] and image processing [4]. However, depending on different factors, e.g., different clustering algorithms, initializations, parameter settings (the number of clusters c), many alternative candidate partitions might be discovered for a fixed dataset.

Cluster validity indices (CVIs) are used to quantify the goodness of a partition. Many CVIs have been proposed and successfully used for this task [5–8]. These measures can be generally divided into two major types: internal and external. If the data are *labeled*, the ground truth partition can be used with an external CVI to explore the match between candidate and ground truth partitions. Since the labeled data may not correspond to clusters proposed by any algorithm, we will refer groups in the ground truth as *subsets*, and algorithmically proposed groups as *clusters*. When the data are *unlabeled* (the real case), an important post-clustering question is how to evaluate different candidate partitions. This job falls to the internal CVIs. One of the most important uses of the external CVIs is to evaluate the comparative quality of internal CVIs on labeled data [9], so that in the real case, some confidence can be placed in a chosen internal CVI to guide us towards realistic clusters found in unlabeled data. *This article is focused on external CVIs.*

External CVIs (or comparison measures), are often interpreted as similarity (or dissimilarity) measures between the ground truth and

* Corresponding author.
E-mail addresses: yalei@student.unimelb.edu.au (Y. Lei), jbezdek@unimelb.edu.au (J.C. Bezdek), simone.romano@unimelb.edu.au (S. Romano), vinh.nguyen@unimelb.edu.au (N.X. Vinh), jeffrey.chan@rmit.edu.au (J. Chan), baileyj@unimelb.edu.au (J. Bailey).

http://dx.doi.org/10.1016/j.patcog.2016.12.003
Received 18 June 2016; Received in revised form 15 October 2016; Accepted 4 December 2016
Available online 08 December 2016
0031-3203/

candidate partitions. The ground truth partition, which is usually generated by an expert in the data domain, identifies the primary substructure of interest to the expert. This partition provides a benchmark for comparison with candidate partitions. The general idea of this evaluation methodology is that the more similar a candidate is to the ground truth (a larger value for the similarity measure), the better this partition approximates the labeled structure in the data.

However, this evaluation methodology implicitly assumes that the similarity measure works correctly, i.e., that a larger similarity score indicates a partition that is really more similar to the ground truth. But this assumption may not always hold. When this assumption is false, the evaluation results will be misleading. One of the reasons that can cause the assumption to be false is that a measure may have bias issues. That is, some measures are biased towards certain clusterings, even though they are not more similar to the ground truth compared to the other candidate partitions being evaluated. This can cause misleading results for users employing these biased measures. Thus, recognizing and understanding the bias behavior of the CVIs is crucial.

The *Rand Index* (RI, similarity measure) is a very popular pair-counting based validation measure that has been widely used in many applications [10–16] in the last five years. It has been noticed that the RI tends to favor candidate partitions with larger numbers of clusters when the number of subsets in the ground truth is fixed [5], i.e., it tends to increase as the number of clusters increases (we call it NCinc bias in this work, where NC=number of clusters). NC bias means that the CVI's preference is influenced by the number of clusters in the candidate partitions. For example, some measures may prefer the partition with larger (smaller) number of clusters, i.e., NCinc (NCdec) bias. The following initial example illustrates NC bias for two popular measures, the *Rand Index* (RI) and *Jaccard Index* (JI) measures.

1.1. Example 1 – NC bias of RI and JI

In this example, we illustrate NC bias for RI and JI. We generate a set of candidate partitions randomly with different numbers of clusters and a random ground truth. We use RI and JI to choose the most similar partition from the candidate partitions by comparing the similarity between each of them and the ground truth. As there is no difference in the generation methodology of the candidate partitions, we expect them to be treated equally on average. A measure without NC bias should treat these candidate partitions equally without preference to any partition in terms of their different number of clusters. However, if a measure prefers the partition, e.g., with a larger number of clusters (gives higher value to the partition with a larger number of clusters if it is a similarity measure), we say it possess NC bias, more specifically, NCinc bias.

Let U_{GT} be a ground truth partition with c_{true} subsets. Consider a set of N=100,000 objects, let the number of clusters in the candidate partitions c vary from 2 to c_{max}, where $c_{max} = 3{*}c_{true}$. We randomly generate a ground truth partition U_{GT} with $c_{true} = 5$. Then for each c, $2 \le c \le 15$, we generate 100 partitions randomly, and calculate the RI and JI between U_{GT} and each generated partition. Finally, we compute the average values of these two measures at each value of c. The results are shown in Fig. 1. Please note that the RI and JI are max-optimal (larger value is preferred). Evidently RI monotonically increases and JI monotonically decreases as c increases. Fig. 1 shows that for this experiment, the RI points to c=15, its maximum over the range of c; and the JI points to c=2, its maximum over the range of c. Both indices exhibit NC bias (RI shows NCinc bias and JI shows NCdec bias).

But, does the RI *always* exhibit NCinc bias towards clusterings with a larger numbers of clusters? *The answer is no*. We have discovered that the overall bias of some CVIs, including the RI, may change their NC bias tendencies depending on the distribution of the subsets in the ground truth. The change in the NC bias status of an external CVI due to the different ground truths is called *GT bias*. This kind of changeable bias behavior caused by the ground truth has not been recognized previously in the literature. It is important to be aware of this phenomenon, since it affects how a user should interpret clustering validation results. Next, we give an example of GT bias (GT=ground truth).

1.2. Example 2 – GT bias of RI

We use the same protocols as in Example 1, but change the distribution of the subsets in the ground truth by randomly assigning 80% of the objects to the first cluster and then randomly assigning the remaining 20% of the labels to the other four clusters for c = 2, 3, 4, 5. Thus, the distribution of the ground truth is heavily skewed (non-uniform). The average values of RI and JI are shown in Fig. 2. The shape of JI in Figs. 1 and 2b is same: it still decreases monotonically with c, exhibiting NCdec bias, and indicating c=2 as its preferred choice. Turning now to the RI, we see that trend seen in Fig. 1a is reversed. The RI in Fig. 2a is maximum at c=2, and *decreases* monotonically as c increases. So the NC bias of RI has changed from NCinc bias to NCdec bias. Thus, RI shows GT bias. To summarize, Examples 1 and 2 show that NC bias is possessed by some external CVIs due to monotonic tendencies of the underlying mathematical model. But beyond this, some external CVIs can be influenced by GT bias, which is due to the way the distribution of the ground truth interacts with the elements of the CVI.

The objective of this article is to study the empirical and theoretical implications of GT bias. To the best of our knowledge, this is the first extensive study of this property for external cluster validity indices. In this work, our contributions can be summarized as follows:

1. We identify the GT bias effect for external validation measures, and also explain its importance.
2. We test and discuss NC bias for 26 popular pair-counting based external validation measures.
3. We prove that RI and related 4 indices suffer from GT bias. And also provide theoretical explanations for understanding why GT bias happens and when it happens on RI and related 4 indices.
4. We present experimental results that support our analysis.
5. We present an empirical example to show that *Adjusted Rand index* (ARI) also suffers from a modified GT bias.

The remainder of the paper is organized as follows. In Section 2 we discuss work related to the bias problems of some external validation measures. We introduce relevant notations and definitions of NC bias and GT bias in Section 3. In Section 4, we briefly introduce some background knowledge about 26 pair-counting based external validation measures. In Section 5, we test the influence of NC bias and GT bias for these 26 measures. Theoretical analysis of GT bias on the RI is presented in Section 6. An experimental example, showing that ARI has GT bias in certain scenarios, is presented in Section 7. The paper is concluded in Section 8.

2. Related work

Several works have discussed the bias behavior of external CVIs. As the conditions imposed on the discussion of the biased behavior are varied, here we classify these conditions into three categories for convenience of discussion: i) general bias; ii) NC bias; iii) GT bias.

General Bias. It has been noticed that the RI exhibits a monotonic trend as both the number of subsets in the ground truth and the number of clusters in the candidate partitions increases [17–19]. However, in our case, we consider the monotonic bias behavior of an external CVI as a function of the number of clusters in the candidate partitions when the number of subsets in the ground truth is fixed.

Wu et al. [20] observed that some external CVIs were unduly influenced by the well known tendency of k-means to equalize cluster sizes. They noted that certain CVIs tended to prefer approximately