1 Vinh, Epps, and Bailey: Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance

1.1 Introduction

When creating ensemble models for clustering, there needs to be a way to compare the results of all the different models in the ensemble. Some measure is needed to quantify how much information is shared between each of the different clusterings. While it may be tempting to use an easily observable variable like the number of shared pairs between clusters as a basis for extracting some consensus clustering, *information theoretic* measures are a more mathematically solid basis of measurement. This paper advocates for the use of the **normalized information distance (NID)** as general purposes measure for comparing clusterings.

1.2 A Review of Cluster Comparison Methods

There are three primary types of cluster comparison methods:

1. Pair Counting based measures For a set S with N items, let $\mathbf{U}$ denote a set of R clusters of S and $\mathbf{V}$ denote some other set of C clusters for S. The overlapping information between $\mathbf{U}$ and $\mathbf{V}$ can be represented in an $R \times C$ contingency table $M = [n_{ij}]_{j=1\rightarrow C}^{i=1\rightarrow R}$ where n_{ij} denotes the number of mutual objects in U_i and V_j. The $\binom{N}{2}$ pairs of the items in S can take one of four values:

   - N_{11}: the number of pairs that are in the same clusters in both $\mathbf{U}$ and $\mathbf{V}$
   - N_{00}: the number of pairs that are in different clusters in $\mathbf{U}$ and $\mathbf{V}$
   - N_{10}: the number of pairs that are in the same clusters in $\mathbf{U}$ but in different clusters in $\mathbf{V}$
   - N_{01}: the number of pairs that are in the same clusters in $\mathbf{V}$ but in different clusters in $\mathbf{U}$

   From this, one can compute a measure of similarity in the form of a *Rand Index* or an *Adjusted Rand Index(ARI)*. However, the ARI which takes values between 0 and 1 is not a proper metric.

2. Set matching based measures In things like classification learning, the classification error rate, based on finding matches between clusters in the two clustering, is a popular measure. However, consider that the number of clusters between two clusterings could be different. If this is the case, then we have to leave out analyzing whole globs of nodes that belonging to

some additional cluster. Furthermore, even when the numbers of clusters are the same, this method does not consider the unmatched part of each cluster pair.

3. Information theoretic based measures: Using concepts from information theory, one can define the relationship between **U** and **V** in terms of their entropies, joint entropy, conditional entropies, and mutual information using the marginal and joint distributions of data items in **U** and **V**. Let $H(U)$ denote some important property of **U**. Let $H(U|V)$ denote how much we know about this important property of **U** given that we already know **V**. Define the mutual information between **U** and **V** to thus be $I(U,V)$, where $I(U,V) = H(U) - H(U|V)$ or equivalently $I(U,V) = H(V) - H(V|U)$. The higher the MI, the more useful the information in **V** helps to predict the cluster labels in **U** and vice-versa.

When picking a clustering comparison measure, one typically would want such a metric to have some of the following desirable properties:

- Has a proper metric property: has the properties of a true metric, namely positive definiteness, symmetry, and the triangle inequality. This allows one to apply many nice theoretical results about metric spaces.

- Normalization: the measure can be adjusted so that it lies within a fixed range. This property is often satisfied in the existing literature, as normalized measures are much better behaved.

- Constant baseline: for two independent clusters, the similarity metric between them should be some constant and low value, ideally zero. The Rand Index and ARI do not exhibit this property.

1.3 Mathematical Properties of Information Theoretic Based Measures

1.3.1 Similarity measures

The mutual information metric, $I(U,V)$ is bounded above by the joint information of all clusters, $H(U,V)$ as well as the information available in any full cluster $min(H(U), H(V))$.

1.3.2 Distance measures

Using these bounds, one can define different measures of distance by subtracting $I(U,V)$ from some known upper bound of the mutual information. Denote these measures with D. One can normalize this measure by dividing the mutual information by some known upper bound and then subtracting this term from 1. Denote these measures with d. One can refer to the table for all 5 of these measures, D_{joint}, D_{max}, D_{sum}, D_{sqrt}, and D_{min}. Of these, it can be shown that D_{sqrt} and D_{min} are not metrics. The normalized variation of information, d_{joint} is a metric, and so is the normalized information distance, d_{max}.

Thus, one can see that d_{joint} and d_{max} are the ideal candidates for being a measure of clustering consensus.

1.4 Adjustment for Chance

Information theoretic measures do not exhibit the desirable constant baseline property. Optimizing based on these procedures tends to detect more clusters than there may be in the ”ground-truth” community structure. This emphasizes the need for an ”adjusted-for-chance” version of the mutual information metrics.

To do this, use a permutation model to calculate the expected mutual information between any two clusterings, and then factor this expected MI into the calculation of your adjusted mutual information. Simply put,

$$AMI(U,V) = \frac{d_{max} - E(d_{max})}{1 - E(d_{max})}$$

If two clusters have normalized mutual information equal to the expected value of such information due to complete randomness, the AMI takes a value of zero. If the two clusters are identical, the AMI takes a values of 1.

Unfortunately, none of the adjusted similarity measures are metrics. For larger sets, the expected mutual information tends toward zero, and thus the adjusted measures tend toward the normalized measures.

Thus, AMI is most useful when you have a small number of objects and a large number of potential clusters.