Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?

Nguyen Xuan Vinh N.X.VINH@UNSW.EDU.AU
Julien Epps J.EPPS@UNSW.EDU.AU
The University of New South Wales, Sydney, Australia & ATP Laboratory, National ICT Australia (NICTA)

James Bailey JBAILEY@CSSE.UNIMELB.EDU.AU
The University of Melbourne, Vic. 3010, Australia & Victoria Research Laboratory, National ICT Australia

Abstract

Information theoretic based measures form a fundamental class of similarity measures for comparing clusterings, beside the class of pair-counting based and set-matching based measures. In this paper, we discuss the necessity of correction for chance for information theoretic based measures for clusterings comparison. We observe that the baseline for such measures, *i.e.* average value between random partitions of a data set, does not take on a constant value, and tends to have larger variation when the ratio between the number of data points and the number of clusters is small. This effect is similar in some other non-information theoretic based measures such as the well-known Rand Index. Assuming a hypergeometric model of randomness, we derive the analytical formula for the expected mutual information value between a pair of clusterings, and then propose the adjusted version for several popular information theoretic based measures. Some examples are given to demonstrate the need and usefulness of the adjusted measures.

1. Introduction

Clustering is the "art" of dividing data points in a data set into meaningful groups. The usefulness of this technique has been proven through its widespread application in virtually all fields of science. Over the past few decades, there have been thousands of papers proposing hundreds of clustering algorithms. The endeavor to find better clustering methods will be hardly feasible without the development of effective measures for clusterings comparison, an open research area which has also received much attention. Various clustering comparison measures have been proposed: besides the class of *pair-counting based measures* including the well-known Adjusted Rand Index (Hubert & Arabie, 1985), and *set-matching based measures*, such as the $\mathcal{H}$ criterion (Meilă, 2005), *information theoretic based measures*, such as the Mutual Information (Strehl & Ghosh, 2002) and the Variation of Information (Meilă, 2005), form another fundamental class of clustering comparison measures.

In this paper, we aim to improve the usability of the class of information theoretic-based measures for comparing clusterings. We first observe that such measures either do not have a fixed bound, or do not have a constant baseline value, *i.e.* average value between random partitions of a data set. Since a measure is meant to provide a comparison mechanism, it is generally preferable that it lies within a predetermined range and has a constant baseline value, so as to facilitate comparison and enhance intuitiveness. For information theoretic-based measures, the former has often previously been accomplished through a normalization scheme, *e.g.* division by the maximum value of the index, while the latter, *i.e.* baseline adjustment, to our knowledge, has not been addressed. As will be seen shortly, unadjusted information theoretic based measures have a considerable inherent bias attributable solely to chance, which potentially reduces their usefulness in a number of common situations, such as measuring the distance from a set of clusterings with different number of clusters to a "true" clustering. In this paper, by assuming a hypergeometric model of randomness, we derive the analytical formula for the expected mutual information value between a pair of clusterings, and then propose the adjusted version for

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

several popular information theoretic based measures. While the assumption of a well specified randomness model is needed for theoretical analysis, our experimental results suggest that the adjusted measures also work well in more realistic scenarios where such an assumption is usually violated.

The paper begins by reviewing some well-known clustering comparison measures in section 2, and discussing the need for baseline adjustment for the class of information theoretic based measures in section 3. The derivation of the adjusted measures is presented in section 4. Selected demonstrations for the new measures are given in section 5 while section 6 concludes the paper.

2. Background and Related Work

Let S be a set of N data points $\{s_1, s_2, \dots s_N\}$. We consider the case of hard clustering. Given two clusterings of S, namely $\mathbf{U} = \{U_1, U_2, \dots, U_R\}$ with R clusters, and $\mathbf{V} = \{V_1, V_2, \dots, V_C\}$ with C clusters ($\cap_{i=1}^{R} U_i = \cap_{j=1}^{C} V_j = \emptyset$, $\cup_{i=1}^{R} U_i = \cup_{j=1}^{C} V_j = S$), the information on cluster overlap between $\mathbf{U}$ and $\mathbf{V}$ can be summarized in the form of a $R \times C$ contingency table $M = [n_{ij}]_{j=1\dots C}^{i=1\dots R}$ as illustrated in Table 1, where n_{ij} denotes the number of objects that are common to clusters U_i and V_j. Based on this contingency table, various cluster similarity indices can be built.

Table 1. The Contingency Table, $n_{ij} = |U_i \cap V_j|$

$\mathbf{U}/\mathbf{V}$	V_1	V_2	$\dots$	V_C	Sums
U_1	n_{11}	n_{12}	$\dots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\dots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\dots$	n_{RC}	a_R
Sums	b_1	b_2	$\dots$	b_C	$\sum_{ij} n_{ij} = N$

2.1. Indices based on Pair Counting

An important class of criteria for comparing clusterings is based upon counting the pairs of points on which two clusterings agree or disagree. Any pair of data points from the total of $\binom{N}{2}$ distinct pairs in S falls into one of the following 4 categories: (1) N_{11}: the number of pairs that are in the same cluster in both $\mathbf{U}$ and $\mathbf{V}$; (2) N_{00}: the number of pairs that are in different clusters in both $\mathbf{U}$ and $\mathbf{V}$; (3) N_{01}: the number of pairs that are in the same cluster in $\mathbf{U}$ but in different clusters in $\mathbf{V}$; (4) N_{10}: the number of pairs that are in different clusters in $\mathbf{U}$ but in the same cluster in $\mathbf{V}$. Explicit formulae for calculating the number of the four types can be constructed using entries in the contingency table (Hubert & Arabie, 1985), *e.g.* $N_{11} = \frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(n_{ij}-1)$. Intuitively, N_{11} and N_{00} can be used as indicators of agreement between $\mathbf{U}$ and $\mathbf{V}$, while N_{01} and N_{10} can be used as disagreement indicators. A well known index of this class is the Rand Index (Rand, 1971), defined straightforwardly as:

$$RI(\mathbf{U}, \mathbf{V}) = (N_{00} + N_{11}) / \binom{N}{2} \tag{1}$$

The Rand Index lies between 0 and 1. It takes the value of 1 when the two clusterings are identical, and 0 when no pair of points appear either in the same cluster or in different clusters in both clusterings, *i.e.* $N_{00} = N_{11} = 0$. This happens only when one clustering consists of a single cluster while the other consists only of clusters containing single points. However this scenario is quite extreme and has little practical value. In fact, it is desirable for the similarity index between two random partitions to take values close to zero, or at least a constant value. The problem with the Rand index is that its expected value between two random partitions does not even take a constant value. Hubert and Arabie (1985), by taking the generalized hypergeometric distribution as the model of randomness, *i.e.* the two partitions are picked at random subject to having the original number of classes and objects in each, found the expected value for $(N_{00} + N_{11})$. They suggested using a corrected version of the Rand index of the form:

$$Adjusted\_Index = \frac{Index - Expected\_Index}{Max\_Index - Expected\_Index} \tag{2}$$

thus giving birth to the (Hubert and Arabie) Adjusted Rand Index (ARI):

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{N}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{N}{2}} \tag{3}$$

where n_{ij}'s are entries in the contingency table and a_i, b_j's are its marginal sums. The ARI is bounded above by 1 and takes on the value 0 when the index equals its expected value (under the generalized hypergeometric distribution assumption for randomness). Besides the Adjusted Rand Index, there are many other, possibly less popular, measures in this class. (Albatineh et al., 2006) discussed correction for chance for a comprehensive list of 28 different indices in this class, a number which is large enough to make the task of choosing an appropriate measure difficult and confusing. Their work, and subsequent extension of (Warrens, 2008), however, showed that after correction for chance, many of these measures become equivalent, facilitating the task of choosing a measure.

2.2. Information Theoretic based Indices

Another class of clustering comparison measures, which are information theoretic based, have also been

employed more recently in the clustering literature (Banerjee et al., 2005; Strehl & Ghosh, 2002; Meilă, 2005). Although there is currently no consensus on which is the best measure, information theoretic based measures have received increasing attention for their strong theoretical background. Let us first review some of the very fundamental concepts of information theory (Cover & Thomas, 1991) and then see how those concepts might be used toward assessing clusterings agreement.

Definition 2.1 *The information entropy of a discrete random variable X, that can take on possible values in its domain $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ is defined by:*

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{4}$$

Definition 2.2 *The mutual information between two random variables X and Y with respective domains $\mathcal{X}$ and $\mathcal{Y}$ is defined by:*

$$I(Y, X) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(y, x) \log \frac{p(y, x)}{p(x)p(y)} \tag{5}$$

The mutual information is a symmetric measure that quantifies the mutual dependence between two random variables, or the information that X and Y share. It measures how much knowing one of these variables reduces our uncertainty about the other. This property suggests that the mutual information can be used to measure the information shared by two clusterings, and thus, assess their similarity. For this purpose, we need to put the clusterings in a statistical context. Let us first define the entropy of a clustering $\mathbf{U}$. Suppose that we pick an object at random from S, then the probability that the object falls into cluster U_i is:

$$P(i) = \frac{|U_i|}{N} \tag{6}$$

We define the entropy associated with the clustering $\mathbf{U}$ as:

$$H(\mathbf{U}) = -\sum_{i=1}^{R} P(i) \log P(i) \tag{7}$$

$H(\mathbf{U})$ is non-negative and takes the value 0 only when there is no uncertainty determining an object's cluster membership, *i.e.* there is only one cluster. Similarly, the entropy of the clustering $\mathbf{V}$ can be calculated as $H(\mathbf{V}) = -\sum_{j=1}^{C} P'(j) \log P'(j)$ where $P'(j) = |V_j|/n$. Now we arrive at the Mutual Information (MI) between two clusterings:

$$I(\mathbf{U}, \mathbf{V}) = \sum_{i=1}^{R} \sum_{j=1}^{C} P(i, j) \log \frac{P(i, j)}{P(i)P'(j)} \tag{8}$$

where $P(i, j)$ denotes the probability that a point belongs to cluster U_i in $\mathbf{U}$ and cluster V_j in $\mathbf{V}$:

$$P(i, j) = \frac{|U_i \cap V_j|}{N} \tag{9}$$

MI is a non-negative quantity upper bounded by both the entropies $H(\mathbf{U})$ and $H(\mathbf{V})$, *i.e.* $I(\mathbf{U}, \mathbf{V}) \leq \min\{H(\mathbf{U}), H(\mathbf{V})\}$. It quantifies the information shared by the two clusterings and thus can be employed as a clusterings similarity measure as in (Banerjee et al., 2005). (Meilă, 2005) suggested using the so-called Variation of Information (VI), which she proved to be a true metric on the space of clusterings:

$$VI(\mathbf{U}, \mathbf{V}) = H(\mathbf{U}) + H(\mathbf{V}) - 2I(\mathbf{U}, \mathbf{V}) \tag{10}$$

(Strehl & Ghosh, 2002) on the other hand, employed a normalized version of the Mutual Information defined as :

$$NMI(\mathbf{U}, \mathbf{V}) = \frac{I(\mathbf{U}, \mathbf{V})}{\sqrt{H(\mathbf{U})H(\mathbf{V})}} \tag{11}$$

The VI is lower bounded by 0 (when the two clusterings are identical) and always upper bounded by $\log(N)$, though tighter bounds are achievable depending on the number of clusters (Meilă, 2005). The Normalized Mutual Information (NMI), on the other hand, has a fixed lower bound of 0 and upper bound of 1. It takes the value of 1 when the two clusterings are identical and 0 when the two clusterings are independent, *i.e.* share no information about each other. In the latter case, the contingency table takes the form of the so-called "independence table" where $n_{ij} = |U_i||V_j|/N$ for all i, j. It can be seen that this scenario is quite intuitive (larger clusters are expected to share more data points), and less extreme than the one where the Rand Index takes on a zero value as described earlier, *i.e.* a clustering contains only 1 cluster while the other contains only singleton clusters.

3. Information Theoretic Measures: is a Correction for Chance Necessary?

Depending upon the specific type of application, it might be preferable for a clustering comparison measure to have fixed bounds. This is accomplished with a normalization scheme such as (11) for the Normalized Mutual Information. The baseline value of such a measure attributable solely to chance agreement is also of interest. For information theoretic based measures, the latter has received less attention. Let us consider the following two motivating examples:

1) Example 1 - Distance to a "true" clustering: given a set of N data points and a clustering $\mathbf{U}$ with R clusters being the "true" clustering. Suppose an algorithm

generates a clustering $\mathbf{V}$ with C clusters, another generates a clustering $\mathbf{V}'$ with C' clusters. We need to assess the goodness of the two clustering algorithms, that is, find out whether $\mathbf{V}$ or $\mathbf{V}'$ is closer to the true clustering $\mathbf{U}$. If $C = C'$ then the situation would be quite simple. Since the setting is quite the same for both $\mathbf{V}$ and $\mathbf{V}'$, we expect the comparison to be "fair" under any particular measure. However if $C \neq C'$ the situation would be more complicated. If a measure without fixed bounds, such as the Variation of Information (VI) were employed, its upper bound would not be the same for $VI(\mathbf{U}, \mathbf{V})$ and $VI(\mathbf{U}, \mathbf{V}')$, and the conclusion that a VI value of, says, 0.3, is better than a VI value of 0.5, might be misleading without knowing the respective upper bounds. In this context, it is preferable to employ a normalized measure such as the Normalized Mutual Information (NMI), with fixed bounds 0 and 1.

The NMI however, is not totally problem-free. We construct a small experiment as follows: consider a set of N data points, let the number of clusters vary from 2 to K_{max} and suppose that the true clustering has $K_{true} = [K_{max}/2]$ clusters. Now for each value of K, generate 10,000 random independent clustering solutions (by assigning each data points to a random clusters with equal probability), and calculate the average NMI (of the form given by (11)), Rand Index (RI) and Adjusted Rand Index (ARI) between those clusterings to the true clustering. The results for various combinations of (N, K_{true}) are given in Figure 1. It can be observed that the unadjusted measures such as the RI and NMI have a monotonic increasing pattern as K increases. For the NMI, the variation is more markedly visible at smaller values of the ratio N/K. Thus even by selecting totally at random, a 7-clusters solution would have a greater chance to defeat a 3-clusters solution, although there isn't any difference in the clustering generation methodology. A measure which has been corrected for chance such as the Adjusted Rand Index, on the other hand, has a baseline value always close to zero, and appears not to be biased in favor of any particular value of K. Thus for this example, an adjusted version of the Mutual Information with correction for chance will be necessary for our purpose.

2) Example 2 - Determining the number of clusters via Consensus Clustering: We start by first providing some background on Consensus Clustering. In an era where a huge number of clustering algorithms exist, the Consensus Clustering idea (Monti et al., 2003; Strehl & Ghosh, 2002; Yu et al., 2007) has recently received increasing interest. Consensus Clustering is not just another clustering algorithm: it rather provides a framework for unifying the knowledge obtained from

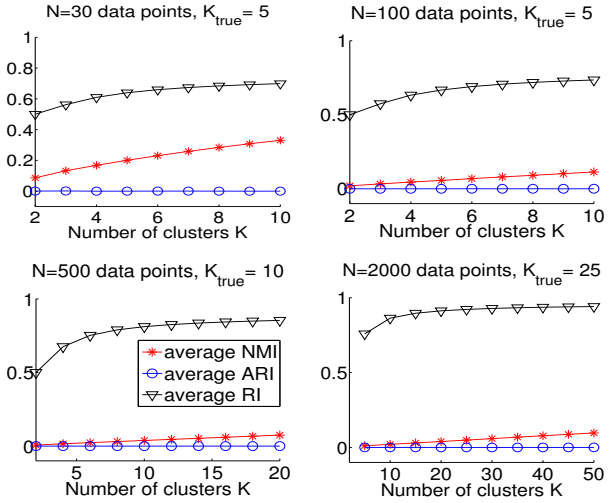

Figure 1. Average distance between sets of random clusterings to a "true" clustering.

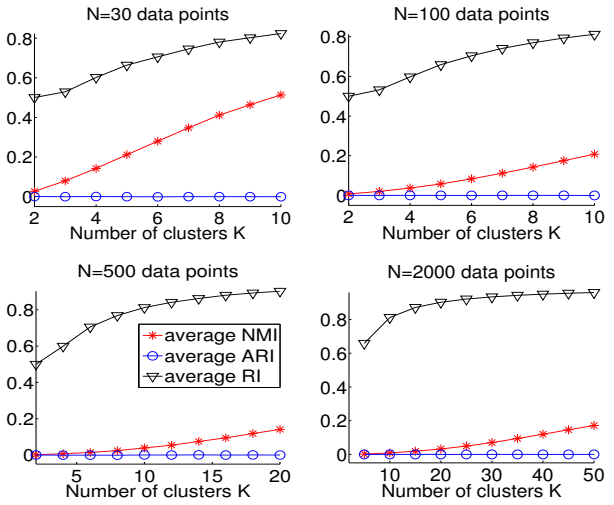

Figure 2. Average pairwise distance within a set of random clusterings, each with the same number of clusters K.

the other algorithms. Given a data set and a single or a set of clustering algorithms, Consensus Clustering employs the clustering algorithm(s) to generate a set of clustering solutions on either the original data set or its perturbed versions. From those clustering solutions, Consensus Clustering aims to choose a robust and high quality representative clustering. Although the main objective of consensus clustering is to discover a high quality cluster structure in a data set, closer inspection of the set of clusterings obtained can often give valuable information about the appropriate number of clusters present. More specifically, we empirically observe that the set of clusterings obtained when the specified number of clusters coincides with the true number of clusters tends to be less diverse, an indication of the robustness of the obtained cluster

structure. To quantify this diversity we have recently developed a novel index (Vinh & Epps, 2009), namely the Consensus Index (CI), which is built upon a suitable clustering similarity measure. Given a value of K, suppose we have generated a set of B clustering solutions $\mathcal{U}_K = \{\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_B\}$, each with K clusters. We define the Consensus Index (CI) of $\mathcal{U}_K$ as:

$$CI(\mathcal{U}_K) = \frac{\sum_{i<j} AM(\mathbf{U}_i, \mathbf{U}_j)}{B(B-1)/2} \tag{12}$$

where the agreement measure AM is a suitable clusterings similarity index. Thus, the Consensus Index CI quantifies the average agreement between all pairs of clustering solutions in a clustering set $\mathcal{U}_K$. The optimal number of clusters K^* is chosen as which that maximizes CI:

$$K^* = \underset{K=2\ldots K_{max}}{\arg\max}\ CI(\mathcal{U}_K) \tag{13}$$

In this setting, it is easily seen that a normalized index is preferable, since the comparison is performed across a wide range of K.

Let us again try the NMI of the form given in (11). We performed a small experiment as follows: given N data points, randomly assign each data point into one of the K clusters with equal probability and check to ensure that the final clustering contains exactly K clusters. Repeat this 200 times to create 200 random clusterings of N data points and K clusters. The average values of NMI, RI and ARI between all 19,900 pairs of clusterings corresponding to this particular value of N and K, *i.e.* averageNMI(N, K), averageRI(N, K) and averageARI(N, K) are recorded. Typical experimental results can be seen in Figure 2. It can be observed that with the same number of data points, the average value of the NMI and RI between random partitions tends to increase as the number of clusters increases, while the average value of the Adjusted Rand Index is always kept very close to zero. When the ratio of N/K is larger, the average value for NMI is reasonably close to zero, but grows as N/K becomes smaller. This is clearly an unwanted effect, since the Consensus Index built upon the NMI would be biased in favour of a larger number of clusters. Thus in this situation, an adjusted version of the MI with correction for chance will be also necessary for our purpose.

4. Correction for Chance

4.1. Model for Randomness

To correct the measures for randomness it is necessary to specify a model according to which random partitions are generated. A common model for randomness is the "permutation model" (Lancaster, 1969, p. 214), in which clusterings are generated randomly subject to having a fixed number of clusters and points in each clusters. This model was adopted by Hubert and Arabie when they derived the adjusted version of the Rand Index. We shall also adopt this model to derive the adjusted version for various information theoretic based measures for comparing clusterings.

Now let us elaborate on the permutation model. Given N data points and two clusterings $\mathbf{U}$ and $\mathbf{V}$ with the number of points in each cluster of the two clusterings fixed, *i.e.* $|U_i| = a_i, |V_j| = b_j, i = 1 \ldots R, j = 1 \ldots C$, then the two marginal sum vectors $a = [a_i]$ and $b = [b_j]$ are constant, satisfying $\sum_{i=1}^{R} a_i = \sum_{j=1}^{C} b_j = N$ (the fixed marginals condition). As all the objects and the clusters are distinguishable, there are

$$\Omega_1 = \binom{N}{a_1}\binom{N-a_1}{a_2}\ldots\binom{N-a_1-\ldots-a_{R-1}}{a_R} \tag{14}$$

different ways to assign the N data points into U_i's, and similarly

$$\Omega_2 = \binom{N}{b_1}\binom{N-b_1}{b_2}\ldots\binom{N-b_1-\ldots-b_{C-1}}{b_C} \tag{15}$$

different ways to assign the N data points into V_j's. Thus the total number of ways to jointly assign N data points into U_i's and V_j's is the product of Ω_1 and Ω_2, which can be simplified as:

$$\Omega = \Omega_1.\Omega_2 = \frac{(N!)^2}{\prod_{i=1}^{R} a_i! \prod_{j=1}^{C} b_j!} \tag{16}$$

Associated with each joint assignment of data points into clusters in the two clusterings is a contingency table $M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}$. Some joint assignments of data points will actually result in identical contingency tables. Indeed, given a (feasible) contingency table M there are:

$$\omega = \frac{N!}{\prod_{i=1}^{R}\prod_{j=1}^{C} n_{ij}!} \tag{17}$$

different ways of assigning the data points into U_i's and V_j's that will actually result in this particular contingency table. It follows that the probability of encountering a particular contingency table from a random clustering formation, subject to the fixed marginals condition is:

$$\mathcal{P}\{M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R} | a, b\} = \frac{\omega}{\Omega} = \frac{\prod_{i=1}^{R} a_i! \prod_{j=1}^{C} b_j!}{N! \prod_{i=1}^{R}\prod_{j=1}^{C} n_{ij}!} \tag{18}$$

Let $\mathcal{M}$ be the set of all the feasible contingency tables M with marginals a and b. The probability distribution of M in $\mathcal{M}$ as specified by (18) is known as the Generalized Hypergeometric distribution (Hubert & Arabie, 1985; Lancaster, 1969).

$$\omega(n_{ij}) = \binom{N}{n_{ij}}\binom{N-n_{ij}}{a_i-n_{ij}}\binom{N-a_i}{b_j-n_{ij}} \prod_{i'=1,i'\neq i}^{R}\binom{N-a_i-\sum_{t=1,t\neq i}^{i'-1} a_t}{a_{i'}} \prod_{j'=1,j'\neq j}^{C}\binom{N-b_j-\sum_{t=1,t\neq j}^{j'-1} b_t}{b_{j'}} \quad (23a)$$

$$E\{I(M)|a,b\} = \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}=(a_i+b_j-N)^+}^{\min(a_i,b_j)} \frac{n_{ij}}{N}\log\left(\frac{N.n_{ij}}{a_i b_j}\right)\frac{a_i!b_j!(N-a_i)!(N-b_j)!}{N!n_{ij}!(a_i-n_{ij})!(b_j-n_{ij})!(N-a_i-b_j+n_{ij})!} \quad (24a)$$

4.2. Expected Mutual Information

In this section we shall calculate the expected value of the Mutual Information between two random clusterings generated by the permutation model described above. Specifically, given two clusterings $\mathbf{U}$ and $\mathbf{V}$ we would like to know the average mutual information between all clustering pairs that have the same number of clusters and data points in each cluster as in $\mathbf{U}$ and $\mathbf{V}$ respectively. The mutual information of such a pair of clusterings can be calculated from the associated contingency table. In fact, let $I(M)$ denote the mutual information between (any) two clusterings associated with the contingency table M. Clearly we have:

$$I(M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b) = \sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N}\log\frac{N.n_{ij}}{a_i b_j} \quad (19)$$

Thus the average mutual information value between all possible pairs of clusterings is actually the expected value of $I(M)$ over the set of the associated contingency tables $\mathcal{M}$. This value is given by:

$$E\{I(M)|a,b\} = \sum_{M\in\mathcal{M}} I(M)\mathcal{P}\{M|a,b\} \quad (20)$$
$$= \sum_{M\in\mathcal{M}}\sum_{i=1}^{R}\sum_{j=1}^{C}\frac{n_{ij}}{N}\log\frac{N.n_{ij}}{a_i b_j}\mathcal{P}\{M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b\}$$

By reordering the sums in (20) we can obtain:

$$E\{I(M)|a,b\} =$$
$$\sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{M\in\mathcal{M}}\frac{n_{ij}}{N}\log\frac{N.n_{ij}}{a_i b_j}\mathcal{P}\{M = [n_{ij}]_{j=1\ldots C}^{i=1\ldots R}|a,b\}$$
$$= \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{n_{ij}}\frac{n_{ij}}{N}\log\frac{N.n_{ij}}{a_i b_j}\mathcal{P}\{M|n_{ij},a,b\} \quad (21)$$

where the sum $\sum_{n_{ij}}(\cdot)$ runs over all possible values of n_{ij}, and $\mathcal{P}\{M|n_{ij},a,b\}$ denotes the probability of obtaining a contingency matrix with marginal sums equal to a and b and the cell at the i-th row and j-th column equals to n_{ij}. Let us now calculate how many different joint assignments of the data points into the two clusterings will result in such a contingency table. We start by picking up the n_{ij} data points that are shared by the clusters U_i and V_j for which there are $\binom{N}{n_{ij}}$ choices. Next we choose the remaining $a_j - n_{ij}$ data points to be assigned to cluster U_i, for which there are $\binom{N-n_{ij}}{a_i-n_{ij}}$ choices. After that, the remaining $b_j - n_{ij}$ data points also need to be chosen for cluster V_j, but those points must not have been chosen for cluster U_i, since U_i and V_j share only exactly n_{ij} points. The number of remaining choices thus reduces to $\binom{N-a_i}{b_j-n_{ij}}$ instead of $\binom{N-n_{ij}}{b_j-n_{ij}}$. After having chosen the data points for U_i and V_j, points can be assigned to all the other clusters in the two clusterings without any restriction. The total number of joint assignments that would result in cluster U_i and V_j sharing exactly n_{ij} data points, denoted by $\omega(n_{ij})$, is therefore given in (23a).

It follows that the probability of encountering a contingency table, with the cell at the i-th row and j-th column equals to n_{ij} from random clusterings is $\mathcal{P}\{M|n_{ij},a,b\} = \omega(n_{ij})/\Omega$ or:

$$\mathcal{P}\{M|n_{ij},a,b\} = \frac{\binom{N}{n_{ij}}\binom{N-n_{ij}}{a_i-n_{ij}}\binom{N-a_i}{b_j-n_{ij}}}{\binom{N}{a_i}\binom{N}{b_j}} \quad (22)$$

Next we investigate the set of feasible values for n_{ij}. Clearly n_{ij} can not exceed $\min(a_i,\, b_j)$. Furthermore, by observing (22) we can see that the value of $N - a_i$ must be larger than or equal to $b_j - n_{ij}$, or equivalently $a_i + b_j - N \leq n_{ij}$, for the validity of the expression $\binom{N-a_i}{b_j-n_{ij}}$. Thus n_{ij} can take on values in the range $[(a_i + b_j - N)^+, \min(a_i, b_j)]$ where $(a_i + b_j - N)^+$ denotes $\max(0, a_i + b_j - N)$. Putting everything together and simplifying, we finally obtain the formula for the expected mutual information given in equation (24a), with the usual conventions $0\log 0 = 1$ and $0! = 1$. Similarly, the expectation for $I^2(M|a,b)$ can be obtained by replacing the term $\frac{n_{ij}}{N}\log(\frac{N.n_{ij}}{a_i b_j})$ in (24a) by $(\frac{n_{ij}}{N}\log(\frac{N.n_{ij}}{a_i b_j}))^2$. The variance of $I(M|a,b)$ can then be calculated as:

$$V\{I(M)|a,b\} = E\{I^2(M)|a,b\} - (E\{I(M)|a,b\})^2 \quad (23)$$

Having calculated the expectation and variance of the mutual information, and assuming that $I(M|a,b)$ has an approximate normal distribution then a one-sided significance test for the value of $I(M|a,b)$ can be built based upon the Z-score:

$$Z = \frac{I(M|a,b) - E\{I(M)|a,b\}}{V\{I(M)|a,b\}} \quad (24)$$

4.3. The Proposed Adjusted Measures

As suggested in (Hubert & Arabie, 1985), the general form of a (similarity) index corrected for chance is

given in (2), which is bounded above by 1 and takes on the value 0 when the index equals its expected value. Having calculated the expectation of the Mutual Information, we propose the Adjusted Mutual Information (AMI) as follows:

$$AMI(\mathbf{U},\mathbf{V}) = \frac{I(\mathbf{U},\mathbf{V}) - E\{I(M)|a,b\}}{\sqrt{H(\mathbf{U})H(\mathbf{V})} - E\{I(M)|a,b\}} \quad (25)$$

where a and b are the marginals of the contingency table of $\mathbf{U}$ and $\mathbf{V}$. Note that it is also possible to define other forms of the AMI, such as:

$$AMI(\mathbf{U},\mathbf{V}) = \frac{I(\mathbf{U},\mathbf{V}) - E\{I(M)|a,b\}}{\max\{H(\mathbf{U}),H(\mathbf{V})\} - E\{I(M)|a,b\}} \quad (26)$$

or:

$$AMI(\mathbf{U},\mathbf{V}) = \frac{I(\mathbf{U},\mathbf{V}) - E\{I(M)|a,b\}}{\frac{1}{2}(H(\mathbf{U}) + H(\mathbf{V})) - E\{I(M)|a,b\}} \quad (27)$$

since $\sqrt{H(\mathbf{U})H(\mathbf{V})}$, $\max\{H(\mathbf{U}),H(\mathbf{V})\}$ and $\frac{1}{2}(H(\mathbf{U}) + H(\mathbf{V}))$ are all valid upper bounds of the Mutual Information. It is interesting to note that the adjusted-for-chance forms of the Mutual Information are all normalized in a stochastic sense. Specifically, the AMI takes a value of 1 when the two clusterings are identical, and 0 when the mutual information between the two clusterings equals its expected value.

For the Variation of Information (VI), since this is a distance measure, an adjustment can be made based on the general formulation (Hubert & Arabie, 1985):

$$Adjusted\_Index = \frac{Expected\_Index - Index}{Expected\_Index - Min\_Index}$$

Therefore the Adjusted Variation of Information (AVI) is given by:

$$AVI(\mathbf{U},\mathbf{V}) = \frac{2I(\mathbf{U},\mathbf{V}) - 2E\{I(M)|a,b\})}{H(\mathbf{U}) + H(\mathbf{V}) - 2E\{I(M)|a,b\}} \quad (28)$$

which turns out to be equivalent to the AMI of the form given in (27). One can also think of using $E\{VI(\mathbf{U},\mathbf{V})\} = H(\mathbf{U}) + H(\mathbf{V}) - 2E\{I(M)|a,b\}$ as a stochastic upper bound for the VI, so as to provide a normalization for this distance measure following the form $Normalized\_Index = Index/Max\_Index$. The Normalized Variation of Information (NVI) is given by:

$$NVI(\mathbf{U},\mathbf{V}) = \frac{H(\mathbf{U}) + H(\mathbf{V}) - 2I(\mathbf{U},\mathbf{V})}{H(\mathbf{U}) + H(\mathbf{V}) - 2E\{I(M)|a,b\}} \quad (29)$$

but this in turn can be shown to be equivalent to $1 - AVI(\mathbf{U},\mathbf{V})$. Therefore, if adjustment or normalization were performed on the Variation of Information, we effectively come back to one of the forms of the Adjusted Mutual Information, *i.e.* that given by (27).

5. Experiment

We repeat the experiments corresponding to the two examples in section 3 with the Adjusted Mutual Information (AMI) of the form in (25), the Adjusted Variation of Information (AVI) (or equivalently the AMI of the form in (27)), and the Adjusted Rand Index (ARI). Results are presented in Figure 3 and Figure 4.

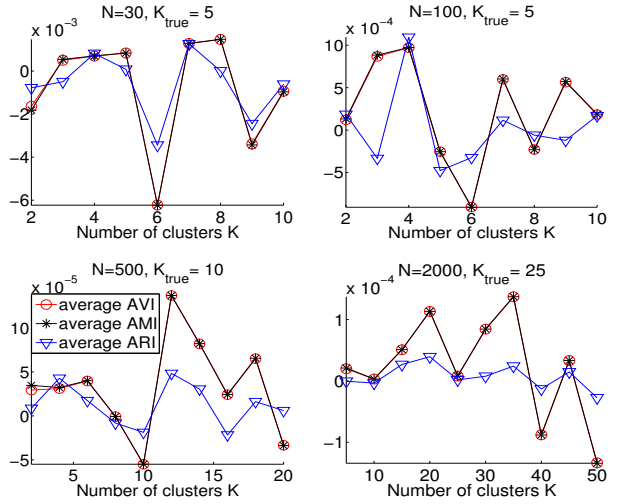

Figure 3. Average adjusted measure values from a set of random clusterings to a fixed, "true" clustering. The values are kept close to zero with negligible variation.

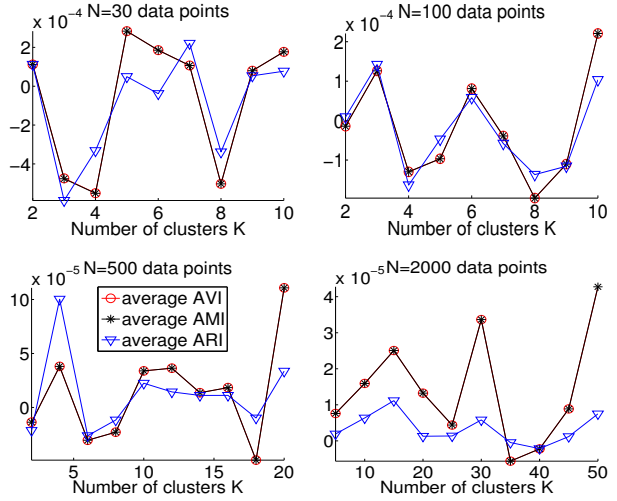

Figure 4. Average pairwise adjusted measure values within sets of random clusterings (each with the same number of clusters K). The values are kept close to zero with negligible variation.

It can be observed that, just like the ARI, the average value of the AMI and AVI between random partitions is now kept close to zero. The difference between the AMI and AVI (which is in fact another form of the

AMI) is hardly discernible in these scenarios. Furthermore, there seems to be a strong correlation between the average values of ARI and AMI. Discovering the subtle similarities and dissimilarities between the AMI and other non-information theoretic based measures such as the ARI is the subject of our another work, of which results are expected to be published in (Vinh et al., 2009).

6. Discussion and Conclusion

In this paper, we have discussed the need for providing correction for chance for some information theoretic based measures for comparing clusterings. Based on the assumption of a hypergeometric model of randomness, we derived the analytical formula for the adjusted measures. We discussed two examples where the adjusted measures are more preferable. Experimental results suggest that the adjusted versions of the information theoretic measures are most useful when the number of data points in each cluster is relatively small, where the variation of the unadjusted measures is markedly recognizable. An example of such a situation is the case of sample clustering for microarray data, where each cluster might contain as few as only 5-7 samples (Monti et al., 2003; Yu et al., 2007).

It should be noted that while the model of randomness assumes that the clusterings must have a fixed number of clusters and fixed number of points in each cluster, the clusterings generated in our experiments do not need to follow such a requirement. The assumptions are needed for the derivation of the analytical results, while in practice, the clusterings generated by clustering algorithms almost never satisfy such assumptions. Nevertheless, the adjusted measures derived under the hypergeometric model of randomness still have a baseline close to zero with negligible variation as observed in various experiments. Although there exist criticisms about the artificiality of the randomness model (Meilă, 2005), in our opinion, the expected value of the measures obtained under such model still conveys more practical information than a theoretical upper bound or lower bound, such as that for the Variation of Information (Meilă, 2005). While the upper/lower bound is generally a single extreme case, the expectation value, on the other hand, tells us on average how a bad clustering, not necessarily very extreme, would score.

Acknowledgement This work was partially supported by the Australian Research Council and partially supported by NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program.

Availability Matlab code for computing the Adjusted Mutual Information (AMI) is available from http://ee.unsw.edu.au/∼nguyenv/Software.htm

References

Albatineh, A. N., Niewiadomska-Bugaj, M., & Mihalko, D. (2006). On similarity indices and correction for chance agreement. *Journal of Classification*, *23*, 301–313.

Banerjee, A., Dhillon, I. S., Ghosh, J., & Sra, S. (2005). Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, *6*, 1345–1382.

Cover, T. M., & Thomas, J. A. (1991). *Elements of information theory*. Wiley.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 193–218.

Lancaster, H. (1969). The chi-squared distribution. . New York: John Wiley.

Meilă, M. (2005). Comparing clusterings: an axiomatic view. *ICML '05: Proceedings of the 22nd international conference on Machine learning* (pp. 577–584). New York, NY, USA: ACM.

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, *52*, 91–118.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*, 846–850.

Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Vinh, N. X., & Epps, J. (2009). A novel approach for automatic number of clusters detection in microarray data based on consensus clustering. *BIBE'09: The IEEE International Conference on BioInformatics and BioEngineering, to appear.*

Vinh, N. X., Epps, J., & Bailey, J. (2009). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *to be submitted.*

Warrens, M. (2008). On similarity coefficients for 2x2 tables and correction for chance. *Psychometrika*, *73*, 487–502.

Yu, Z., Wong, H.-S., & Wang, H. (2007). Graph-based consensus clustering for class discovery from gene expression data. *Bioinformatics*, *23*, 2888–2896.