Cluster Validation Using Mutual Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of evaluating a hard clustering algorithm given ground truth. We accomplish this using the mutual information (MI) between the algorithm's results and a set of "truth" labels. While others have proposed an adjustment to the MI to account for results obtained by chance, the existing methods do not consider that the clustering algorithm result is drawn from a distribution of clustering results while the given truths are fixed. By considering the problem in this way, we develop a distribution of contingency tables with fixed row sums and a fixed number of columns. The distribution involves a ratio of Stirling numbers of the second kind. Using this distribution we derive a new adjusted mutual information measure (AMI*). We also propose a normalized version of our measure. Finally, we demonstrate empirically, that the new measure for scoring clustering results given ground truth generates preferred results over the existing techniques.

1 Introduction

In this paper, we present a method for evaluating results from hard clustering algorithms. Hard clustering algorithms are algorithms which partition a dataset into mutually exclusive and collectively exhaustive subsets. Specifically we are interested in the problem of measuring the quality of a clustering result given "ground truth" labels of the clustered data.

Having a measure of the quality of a clustering result is essential for validating various clustering design choices such as the choice of features, metrics, and clustering algorithm. An ideal measure should be able to compare one clustering result to another result, even when the number of clusters differ from one result to the other. In addition the measure should stand on its own in the sense that one should be able to judge an algorithm's applicability to different data sets using the measure. In other words, the measure should reflect the *clusterability* of one data set compared to another using a particular algorithm.

The general problem of comparing two ways of clustering data has been the subject of much investigation in the past. In fact, a recent survey identified nearly two dozen such techniques[1]. In order for these measures to be broadly applicable, they need to be able to fairly compare clusterings when the number of clusters differs from one clustering to the next. One way to accomplish this is to construct the measure relative to results that we would expect to obtain had clustering been performed at random. This was initially done for the Rand Index[11], a commonly used similarity measure comparing two clusterings, first in 1984 using an asymptotic form of the joint clustering distribution by Morey and Agresti[9]. An exact expression of the joint distribution and the resulting adjusted Rand Index followed soon after in 1985 by Hubert and Arabie [5].

Many of the existing measures are adjusted in a similar way to correct for results obtained by chance. The Hubert and Arabie hypergeometric joint distribution has become the *de facto* standard for making measure adjustments.

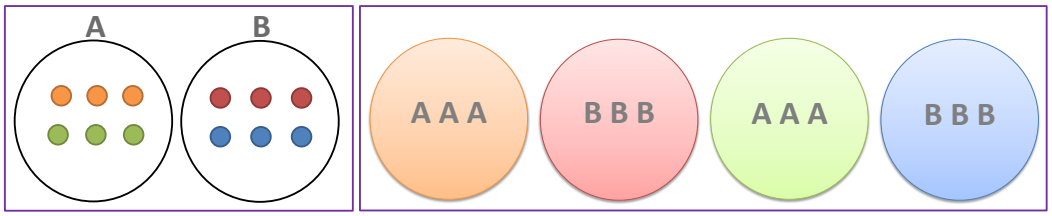


Figure 1: **Two example clusterings** are shown on the left and right. The first example (left) divides objects into cluster A and B and the corresponding ground truth is indicated by the color. In the second example (right), objects are divided into four clusters and the ground truth is labelled as A or B. In these two examples, a symmetric clustering measure would score these two examples the same despite the fact that the second clustering was clearly more successful in recovering the given ground truth.

Meilă[8] axiomatically studied many measures for comparing clusterings. While Meilă's paper asserts that the measurement task is important to measure how close a clustering result is "to a gold standard clustering", the paper then seems to address distance measures between clusterings. In fact, existing measures for comparing two ways of clustering tend to support the assumption that the comparison should be symmetric. As far as we know this assumption has been implicitly repeated since 1984. However, we challenge this assumption in the context of our application where we are comparing a clustering result with respect to some *truth*.

Not only is a symmetric comparison measure unnecessary but also it is undesirable if we want the measure to stand on its own as a measure of clusterability. This is easily demonstrated with the example shown in Figure 1. In this figure we show two clustering examples that are equivalent given a symmetric measure. However, the second clustering is clearly better than the first given the ground truth.

In this paper we introduce a new approach to measure clustering quality given ground truth. We accomplish this by creating an adjusted version of the Mutual Information (MI) measure. We use MI because it identifies the information content of the truth that is captured by the clustering (see [12, 7]). Recently Vihn, Epps and Bailey[14, 15] have derived an adjusted mutual information (AMI) measure which follows from the hypergeometric distribution. In this paper, we derive an adjusted mutual information measure (AMI*) using a more appropriate distribution for cluster validation.

We will show that while our version yields similar results to AMI, the results are not equivalent. In fact, we observe that our measure, AMI*, and the previous version, AMI, seem to differ more often for clusterings that are *far from perfect*. However, we will show that most possible clusterings are *far from perfect*. Additionally, we will demonstrate empirically that when AMI and AMI* disagree, AMI* produces a superior measure of clustering quality.

The rest of this paper is organized as follows: Section 2 reviews contingency tables as a means for tabulating clustering similarity and measuring MI on the same data set. In Section 3 we derive a new baseline adjustment to MI for evaluating clustering quality given ground truth labels. In Section 4 we experimentally compare AMI to AMI*. Finally in Section 5 we summarize our work and future directions.

2 Background

In this section we will describe contingency tables used in the context of our application of describing clustering results given a set of ground truth labels. We then describe mutual information (MI) between the clustering and the ground truth and show how it can be estimated from a contingency table.

2.1 Contingency Tables

Let $S = \{s_1, s_2, ..., s_N\}$ be a set of N objects, where each s_i belongs to one of R true categories labelled $\mathcal{U} = \{u_1, u_2, ..., u_R\}$. A clustering algorithm produces a partition of these N objects into

C clusters labelled $\mathcal{V} = \{v_1, v_2, ..., v_C\}$. We can summarize the overlap between the true categories and the clusters produced by a clustering algorithm, in the form of a contingency table M, where table element M_{ij} is the number of objects belonging to category u_i, that was placed into cluster v_j. We also define $a_i = \sum_{j=1}^{C} M_{ij}$ (row sums) and $b_j = \sum_{i=1}^{R} M_{ij}$ (column sums). This is shown below:

$\mathcal{U}/\mathcal{V}$	v_1	v_2	...	v_C	Sums
u_1	M_{11}	M_{12}	...	M_{1C}	a_1
u_2	M_{21}	M_{22}	...	M_{2C}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
u_R	M_{R1}	M_{R2}	...	M_{RC}	a_R
Sums	b_1	b_2	...	b_C	N

Figure 2: **A Contingency Table** M is shown in the unshaded region of the table above. M_{ij} is the number of objects belonging to category u_i that were placed in cluster v_j. a_i and b_j denote row sums and column sums respectively.

2.2 Mutual Information (MI)

Formally, the mutual information $I(X;Y)$ between discrete random variables X and Y is defined as

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the domains of X and Y respectively. $I(X;Y)$ is a symmetric measure that quantifies the information X and Y share. It is also useful to introduce the concept of entropy, denoted by $H(X)$, which is a measure of uncertainty associated with a random variable, X. Formally,

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{2}$$

It is easy to verify that

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{3}$$

Thus, MI is a measure of how much knowing one of the variables reduces our uncertainty of the other. Also note that $I(X;Y)$ is upper-bounded by both $H(X)$ and $H(Y)$.

2.3 Estimating MI from a contingency table

Adopting a statistical view, we introduce random variables $U \in \mathcal{U}$ to represent the category, and $V \in \mathcal{V}$ to represent the cluster, that an object belongs to. Then after observing a contingency table M, we have the following frequentist estimates:

$$Pr(U = u_i) = \tfrac{a_i}{N}, \qquad Pr(V = v_j) = \tfrac{b_j}{N}, \quad \text{and} \quad Pr(U = u_i, V = v_i) = \tfrac{M_{ij}}{N} \tag{4}$$

Now, we can obtain an estimate for the mutual information between U and V as:

$$\hat{I}(M) = \sum_{i=1}^{R} \sum_{j=1}^{C} \Theta(M_{ij}, a_i, b_j) \quad \text{where,} \quad \Theta(n,a,b) = \frac{n}{N} \log \frac{nN}{ab} \tag{5}$$

From now on, we will use the same variables (e.g U, V) to represent both the random variables representing the category/cluster labels of an object, and the partition of the objects induced by these labels, unless the distinction is not obvious from the context.

3 Adjusted Mutual Information (AMI)

Consider the scenario where we are comparing two partitions, V and V' with C and C' clusters respectively, against a"true" partition U. If $C = C'$, the MI of each of the two partitions to the true partition, $I(U;V)$ and $I(U;V')$, is a fair measure for comparing these clustering algorithms.

However, if $C \neq C'$ this might not be the case. For example, suppose we are comparing 3 partitions, V_1,V_2 and V_3 of a dataset consisting of two objects from one category and two objects from another. Figure 3 shows the contingency tables of V_1,V_2 and V_3 with respect to U, the "true" partition induced by the labels. V_1 is the best possible clustering of the dataset whereas V_3 is a completely uninformative "clustering" since it placed each object in a separate "cluster". But it turns out that the mutual information for all the 3 contingency tables in Figure 3 are the same ($= H(U)$). Also note that, any random partition of N objects into N categories, although completely uninformative, achieves the highest possible Mutual Information score with respect to the true clustering.

2	0
0	2

(a) U vs. V1

2	0	0
0	1	1

(b) U vs. V2

1	1	0	0
0	0	1	1

(c) U vs. V3

Figure 3: **Pitfalls of using MI** to compare algorithms which produce different number of clusters. MI is the same for (a), (b) and (c) despite (a) being a better clustering.

This example suggests that a more informative measure should include a correction term to account for the mutual information that would be obtained by chance. To evaluate an algorithm that partitions the data into C clusters, we must look at how much better this algorithm does, on average, than an algorithm that randomly partitions the same data into C clusters. Vinh et al [14] suggested an expected mutual information (EMI) correction: given two clusterings which partition data points into clusters of size $\mathbf{a} = [a_1, ..., a_R]$ and $\mathbf{b} = [b_1, ..., b_C]$ respectively, they computed EMI over all possible pairs of such clusterings of the data.

Our formulation is as follows: Given the cardinalities of the true clusters, $\mathbf{a} = [a_1, ..., a_R]$, and the number of clusters, C, produced by an algorithm, we calculate the expected mutual information between the true clustering and all possible clusterings of this data into *exactly* C clusters. Note that we do not fix $\mathbf{b} = [b_1, ..., b_C]$, the cardinalities of the clusters V. We will refer to this definition of expected mutual information by EMI* to distinguish this from the EMI defined by Vinh et al [14].

We describe how EMI* is calculated in Section 3.1. In Section 3.2, we briefly explore the relation between EMI and EMI*. Then, in Section 3.3 we use EMI* as a baseline that can be subtracted from MI to obtain a new cluster validation measure (AMI*). We also introduce a normalized version of this measure (NAMI*). Finally, in Section 3.4, we describe an approximation method for computing EMI* for large datasets.

3.1 Expected Mutual Information (EMI*)

Suppose we are given N objects, with $a_i > 0$ objects belonging to categories u_i for $i = 1...R$. Let us compute the expectation of the mutual information estimate over all possible clusterings of these objects into *exactly* C clusters. We have,

$$\text{EMI}^* = \mathbb{E}[\hat{I}(M)|\mathbf{a}, C] = \sum_{M \in \mathcal{M}} \hat{I}(M) P(M|\mathbf{a}, C) \tag{6}$$

where, $\mathcal{M}$ is the set of all $R \times C$ contingency tables M, such that $\sum_{j=1}^{C} M_{ij} = a_i$ (row i sums to a_i) for $i = 1...R$ and $\sum_{i=1}^{R} M_{ij} > 0$ for $j = 1...C$ (columns sums are non zero), and $P(M|\mathbf{a}, C)$ is calculated as:

$$P(M|\mathbf{a}, C) = \frac{\mathcal{N}(M)}{\sum_{M \in \mathcal{M}} \mathcal{N}(M)} \tag{7}$$

where $\mathcal{N}(M)$ is the number of ways to cluster the given objects that result in the contingency table M. Plugging in (5) and (7) in (6), we have,

$$\mathbb{E}[\hat{I}(M)|\mathbf{a}, C] = \sum_{M\in\mathcal{M}} \left\{ \left[\sum_{i=1}^{R}\sum_{j=1}^{C} \Theta(M_{ij}, a_i, b_j) \right] \left[\frac{\mathcal{N}(M)}{\sum_{M\in\mathcal{M}} \mathcal{N}(M)} \right] \right\}$$
$$= \frac{1}{\sum_{M\in\mathcal{M}} \mathcal{N}(M)} \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{M\in\mathcal{M}} [\Theta(M_{ij}, a_i, b_j)\mathcal{N}(M)] \quad (8)$$

The summation over M in (8) can be replaced with a summation over all possible values for b_j and M_{ij}. We need not sum over a_i since it is a fixed quantity. Let us now consider the range of values that b_j and M_{ij} can take. Since there must be at least one element in each column of M, it is easy to see that b_j has to be at least 1 and at most $N-(C-1)$. Given b_j, M_{ij} can be at most $\min(a_i, b_j)$. Also note that, after filling the $[i,j]^{th}$ cell, the j^{th} column must be filled with $b_j - M_{ij}$ elements from a pool of $N-a_i$ elements. Therefore, $M_{i,j}$ has to be at least $(a_i+b-N)^+ \triangleq \max(0, a_i+b_j-N)$.

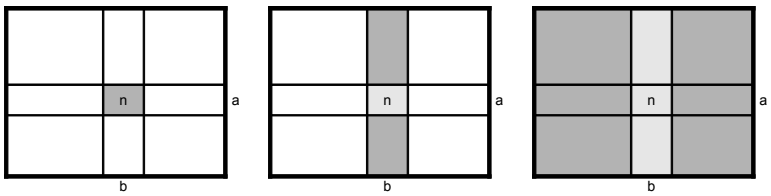
To replace the summation over M in (8) as mentioned above, we also need to replace $\mathcal{N}(M)$ with $\mathcal{N}(M_{ij}, a_i, b_j|C)$, where $\mathcal{N}(n, a, b|C)$ is the number of ways to cluster the given objects into exactly C clusters such that there are n elements in a particular cell, and the number of elements in the corresponding row and column are a and b respectively. With this transformation (8) becomes:

$$\mathbb{E}[\hat{I}(M)|\mathbf{a}, C] = \frac{1}{\sum_{M\in\mathcal{M}} \mathcal{N}(M)} \sum_{i=1}^{R}\sum_{j=1}^{C}\sum_{b=1}^{N-C+1}\sum_{n=(a_i+b-N)^+}^{min(a_i,b)} [\Theta(n, a_i, b)\mathcal{N}(n, a_i, b|C)] \quad (9)$$

Since the categories of the objects are given, $\sum_{M\in\mathcal{M}} \mathcal{N}(M)$ is just the number of ways to partition N distinguishable objects into C distinguishable non-empty bins, i.e:

$$\sum_{M\in\mathcal{M}} \mathcal{N}(M) = \left\{ {N \atop C} \right\} \times C! \quad (10)$$

where $\left\{ {N \atop C} \right\}$ denotes a Stirling numbers of the second kind [3]. We now describe how $\mathcal{N}(n, a, b|C)$



(a) (b) (c)

Figure 4: **Calculating** $\mathcal{N}(n, a, b|C)$, the number of contingency tables with n elements in a particular cell, and a and b elements in the corresponding row and column respectively. First, in (a), the given cell is filled with n elements chosen from a elements in the corresponding row. Then, in (b), the rest of the $b-n$ elements in the corresponding column is filled from a pool $N-a$ elements. Finally, in (c), the remaining $N-b$ elements fill the cells in the remaining $C-1$ columns.

can be calculated (see Figure 4). As mentioned, this is the number of ways to cluster the given N objects into exactly C clusters so that a given cell contains n elements and there are a and b elements in its corresponding row and column respectively. Since there are a elements in the given row, the given cell (darkly shaded in Fig. 4a) can be filled with n elements in $\binom{a}{n}$ ways. After this, the corresponding column needs to be filled with an additional $b-n$ elements (darkly shaded region in Fig. 4b). Since, we cannot use any more elements from the corresponding row to fill up this column, we have only $N-a$ elements available. So, we have to choose $b-n$ elements from a pool of $N-a$ elements, and this can be done in $\binom{N-a}{b-n}$ ways. Finally, there are $N-b$ elements left to fill the cells in the remaining $C-1$ columns of the table (darkly shaded region in Fig. 4c). Since the category of the objects are given, the row to which each element can be assigned is fixed. So, we only need

to consider the number of ways that these $N - b$ elements can be arranged into $C - 1$ columns, so that none of the columns are empty. This is given by $\left\{{N-b \atop C-1}\right\} \times (C-1)!$ and we have,

$$\mathcal{N}(n, a, b|C) = \binom{a}{n}\binom{N-a}{b-n}\left\{{N-b \atop C-1}\right\}(C-1)! \tag{11}$$

Plugging (11) and (10) into (9), we notice that the terms inside the summation are independent of j and hence the summation over j can be removed and the whole expression multiplied by C. Thus, (9) becomes:-

$$\text{EMI}^* = \sum_{i=1}^{R}\sum_{b=1}^{N-C+1}\sum_{n=\max(1,a_i+b-N)}^{\min(a_i,b)} \Theta(n, a_i, b)\frac{\binom{a_i}{n}\binom{N-a_i}{b-n}\left\{{N-b \atop C-1}\right\}}{\left\{{N \atop C}\right\}} \tag{12}$$

3.2 Relation to EMI

To see the relation between EMI and EMI*, note that Eqn. (12) can be written as follows:

$$\text{EMI}^* = \sum_{i=1}^{R}\sum_{j=1}^{C}\left[\sum_{b=1}^{N-C+1}\left\{\sum_{n=\max(1,a_i+b-N)}^{\min(a_i,b)} \Theta(n, a_i, b)\text{p}_n(n; a_i, b, N)\right\}\text{p}_b(b; N, C)\right] \tag{13}$$

where

$$\text{p}_n(n; a, b, N) = \frac{\binom{a}{n}\binom{N-a}{b-n}}{\binom{N}{b}} \qquad \text{and} \qquad \text{p}_b(b; N, C) = \frac{1}{C}\binom{N}{b}\frac{\left\{{N-b \atop C-1}\right\}}{\left\{{N \atop C}\right\}} \tag{14}$$

Here p_n is the density of the hyper-geometric distribution $\mathcal{H}(N, a, b)$ and p_b is the density of an unknown distribution which we will call the "Stirling distribution" $\mathcal{S}(N, C)$, since it involves the Stirling numbers of the second kind [1]. It is easy to see that:

$$\text{EMI}^* = \sum_{i=1}^{R}\sum_{j=1}^{C}\mathbb{E}_{b|N,C}\left[\mathbb{E}_{n|N,a_i,b}\left[\Theta(n, a_i, b)\right]\right] \tag{15}$$

In contrast, the EMI developed by Vinh et al [14] is given by:

$$\text{EMI} = \sum_{i=1}^{R}\sum_{j=1}^{C}\mathbb{E}_{n|N,a_i,b_j}\left[\Theta(n, a_i, b_j)\right] \tag{16}$$

3.3 Normalization

Once we have calculated EMI*, we can calculate the adjusted mutual information as,

$$\text{AMI}^* = \hat{I}(M) - \mathbb{E}[\hat{I}(M)|\mathbf{a}, C] \tag{17}$$

The AMI* measure can be normalized by dividing by some value $\mathcal{K}$ to ensure that it lies within an interpretable range. The normalization schemes suggested by Vinh et al [14] such as $\sqrt{H(U)H(V)} - \text{EMI}$, $\max\{H(U), H(V)\} - \text{EMI}$ or $\frac{1}{2}(H(U) + H(V)) - \text{EMI}$ are not appropriate for our purposes. Since we are using this measure to compare different clustering algorithms, the normalization should remain the same for any clustering of a given dataset. Thus $\mathcal{K}$ should depend only on statistics of the given dataset, and not on the results of any clustering algorithm we are trying to evaluate. We argue that $H(U)$ is a more appropriate normalization constant, since $\hat{I}(M)$ is upper-bounded by $H(U)$, and $H(U)$ is independent of any clustering result. We can also subtract $\mathbb{E}[\hat{I}(M)|\mathbf{a}, R]$ from $H(U)$ so that AMI* is equal to 1 when an algorithm exactly reproduces the true clustering. Thus we have the normalized AMI* given by:

$$\text{NAMI}^* = \frac{\hat{I}(M) - \mathbb{E}[\hat{I}(M)|\mathbf{a}, C]}{H(U) - \mathbb{E}[\hat{I}(M)|\mathbf{a}, R]} \tag{18}$$

[1] Although we call this the Stirling distribution, it should not be confused with other distributions in the literature involving Stirling numbers of the second kind, see e.g. [4, 10]

3.4 Large Sample Approximations

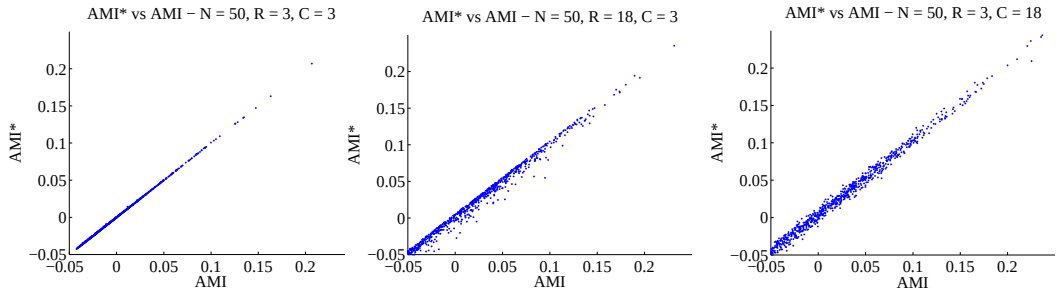
Since the coefficients in (12) involve Stirling numbers of the second kind and factorials which can grow rapidly, computing them directly can cause overflow problems. Although the factorials and Stirling numbers can be very large, each coefficient is a product of two probabilities and is always a quantity between 0 and 1. Therefore, we use an approximate formula to calculate the coefficients for large N.

The approximate form involves approximating the Stirling number of the second kind using an asymptotic form derived by Temme [13]. The solution requires computation of the Lambert W function (a.k.a. the Omega function or Product logarithm) for which we use Halley's iteration method [2].

We can compute the factorials appearing in (12) using Lanczos' approximation for the Gamma Function [6]. In practice, we do not compute the factorials directly but instead work with the logarithm of the factorial $L(x) = ln(x!)$. Using Temme's approximation and $L(x)$, we can easily approximate the coefficients in a manner that avoids overflow issues.

4 Experiments

Ironically, it is difficult to devise a quantitative measure for comparing two measures for comparing clustering algorithms, since any such measure would itself be yet another measure for comparing clustering algorithms! Therefore, we do not attempt to do this, but instead try to illustrate how AMI* differs from AMI.



(a) (b) (c)

Figure 5: **Correlation between AMI* and AMI:** Scatter plots of AMI vs. AMI* for different values of R and C with N = 50.

We first studied the correlation between AMI and AMI* at different values of N, R and C. We randomly generated a large number of contingency tables with a fixed value of N, R and C and compared the AMI and AMI* values for each of these tables. We found that the correlation between AMI and AMI* is high when R and C are low compared to N, but decreases when either R or C is increased. This phenomenon is illustrated in Figure 5, where we have shown scatter plots of AMI versus AMI* at different values or R and C for fixed N.

We also looked at the number of pairs of contingency tables which are ordered differently by AMI and AMI*. We observed that the two measures disagree more often when the AMI/AMI* values are low, i.e. when the corresponding clusterings are far from perfect. In fact, most possible clusterings are far from perfect (e.g. see Figure 6) and it is essential that a measure performs well on these cases.

Since there is no quantitative measure for comparing cluster comparison measures, we performed a qualitative analysis by generating sets of clusterings which are ordered differently by the two measures. We asked colleagues at our institution to order the clusterings based on which clustering they thought seemed better given the truth. We then compared the results to those produced by AMI and AMI*. From the survey results, we tested the hypothesis that AMI* is a superior measure for comparing clusterings, and obtained a p-value of 0.012 from 51 experiments. An example test set is

shown in Figure 7. For this particular example, 94% of the people ranked A over C, 65% ranked B over C and 75% ranked A over B.

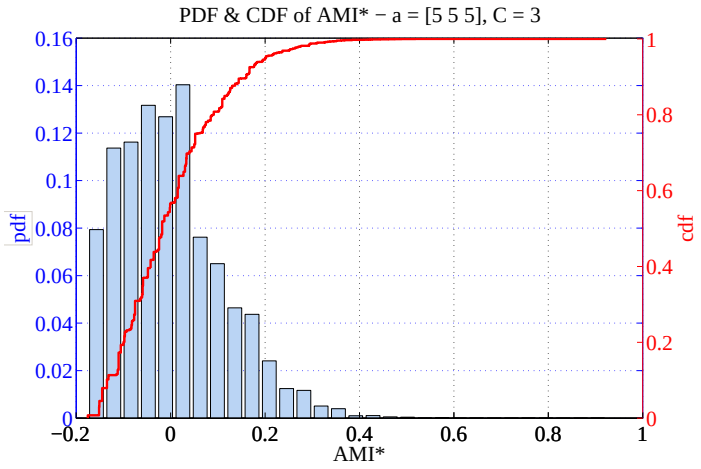


Figure 6: **Distribution and density of AMI*:** We plot the empirical pdf and cdf of AMI* computed by generating all possible clusterings of a dataset consisting of 15 objects into 3 clusters. The dataset contains 5 objects each from 3 true categories. Note that more than 50% of possible clusterings have an AMI* value less than 0, and close to 95% have an AMI* value less than 0.2 (far from perfect).

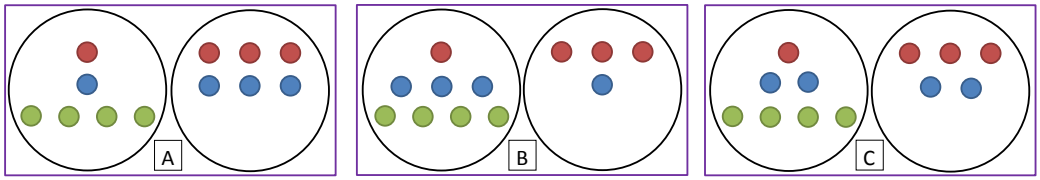


Figure 7: **An example where AMI and AMI* disagree:** Three clusterings of a given dataset into two clusters are shown. The dataset consists of 4 objects each from 3 true categories (indicated by color). Both AMI and AMI* agree that the clustering A is superior to the other two. However, AMI prefers C over B whereas AMI* prefers B over C.

5 Conclusion

In this paper we introduced a new method for evaluating clustering results given ground truth. It is our opinion that our approach addresses a long standing (27 years) flaw in the application of many cluster validation methods.

We described how our approach can be applied to the Mutual Information measure and we experimentally demonstrated that our method generates more intuitive results.

We also described a normalized version of our AMI* measure. The normalization attempts to set a bound on the normalized measure so that the measure can be used across data sets. However, a proof of the maximum AMI* value has thus far been elusive and still requires further investigation.

Our future work will involve applications of this new theory to various machine learning problems.

References

[1] A.N. Albatineh, M. Niewiadomska-Bugaj, and D. Mihalko. On similarity indices and correction for chance agreement. *Journal of Classification*, 23(2):301–313, 2006.

[2] R.M. Corless, G.H. Gonnet, D.E.G. Hare, D.J. Jeffrey, and D.E. Knuth. On the LambertW function. *Advances in Computational mathematics*, 5(1):329–359, 1996.

[3] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete mathematics: a foundation for computer science*, volume 2. Addison-Wesley, 1994.

[4] F. Hennecart. Stirling distributions and stirling numbers of the second kind. computational problems in statistics. *Kybernetika*, 30(3):279–288, 1994.

[5] L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2(1):193–218, 1985.

[6] C. Lanczos. A precision approximation of the gamma function. *Journal of the Society for Industrial and Applied Mathematics: Series B, Numerical Analysis*, 1:86–96, 1964.

[7] M. Meilă. Comparing clusterings by the variation of information. In *Learning theory and Kernel machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003: proceedings*, volume 2777, page 173. Springer Verlag, 2003.

[8] M. Meilă. Comparing clusterings: an axiomatic view. In *Proceedings of the 22nd international conference on Machine learning*, pages 577–584. ACM, 2005.

[9] L.C. Morey and A. Agresti. The measurement of classification agreement: an adjustment to the rand statistic for chance agreement. *Educational and Psychological Measurement*, 44(1):33, 1984.

[10] SB Nandi and SK Dutta. Some discrete distributions involving stirling numbers. *Sankhyā: The Indian Journal of Statistics, Series B*, 48(3):301–314, 1986.

[11] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, pages 846–850, 1971.

[12] A. Strehl and J. Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

[13] N.M. Temme. Asymptotic estimates of Stirling numbers. *Stud. Appl. Math*, 89(3):233–243, 1993.

[14] N.X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1073–1080. ACM, 2009.

[15] N.X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 9999:2837–2854, 2010.