Standardized Mutual Information for Clustering Comparisons: One Step Further in Adjustment for Chance

Simone Romano SIMONE.ROMANO@UNIMELB.EDU.AU
James Bailey BAILEYJ@UNIMELB.EDU.AU
Nguyen Xuan Vinh VINH.NGUYEN@UNIMELB.EDU.AU
Karin Verspoor KARIN.VERSPOOR@UNIMELB.EDU.AU
Department of Computing and Information Systems, The University of Melbourne, Victoria, Australia

Abstract

Mutual information is a very popular measure for comparing clusterings. Previous work has shown that it is beneficial to make an adjustment for chance to this measure, by subtracting an expected value and normalizing via an upper bound. This yields the constant baseline property that enhances intuitiveness. In this paper, we argue that a further type of statistical adjustment for the mutual information is also beneficial – an adjustment to correct selection bias. This type of adjustment is useful when carrying out many clustering comparisons, to select one or more preferred clusterings. It reduces the tendency for the mutual information to choose clustering solutions i) with more clusters, or ii) induced on fewer data points, when compared to a reference one. We term our new adjusted measure the *standardized mutual information*. It requires computation of the variance of mutual information under a hypergeometric model of randomness, which is technically challenging. We derive an analytical formula for this variance and analyze its complexity. We then experimentally assess how our new measure can address selection bias and also increase interpretability. We recommend using the standardized mutual information when making multiple clustering comparisons in situations where the number of records is small compared to the number of clusters considered.

1. Introduction

Clustering techniques aim at partitioning data by grouping objects with similar characteristics into homogeneous clusters (Aggarwal & Reddy, 2013). External validation techniques, which compare a clustering solution to a reference clustering, are therefore extremely important to assess clustering quality. However, there is no acknowledged measure of choice to compare partitions and in practice many measures are used (Wu et al., 2009). Among those, there exist pair-counting based measures such as the Rand Index (RI) (Rand, 1971) and Jaccard coefficient (Ben-Hur et al., 2001), as well as information theoretic measures such as the Mutual Information (MI) (Cover & Thomas, 2012) and Variation of Information (VI) (Meilă, 2007). We provide a glossary of acronyms used in this paper in Table 1.

A desirable property of clustering comparison measures is to have a constant baseline in the case of random independent partitions. Adopting a probabilistic interpretation of the partition problem, an expected value can be computed under the assumption of random and independent clusterings and then subtracted from the measure. Adjusted for chance measures include the Adjusted Rand Index (ARI) (Hubert & Arabie, 1985) and the Adjusted Mutual Information (AMI) (Vinh et al., 2009). They take a 0 expectation value when partitions are independent, and are bounded above by 1 via the use of a normalization factor (an upper bound of the measure).

In this paper, our focus is on information theoretic measures, in particular the mutual information. Our first key observation is that employing a baseline adjustment to the mutual information does not guarantee that it is bias free. In fact, it is still susceptible to what we term *selection bias*, a tendency to choose inappropriate clustering solutions i) with more clusters, or ii) induced on fewer data points, when compared to a reference one. To illustrate, consider the following experiment on a dataset of 500 records: a reference clustering of 10 equal-size clusters is compared in turn with 6 clustering solutions. Each solution is randomly generated with equal-size clusters and the number of clusters in each is 2, 6, 10, 14, 18 and 22 respectively. We then use the AMI measure to select the most similar clustering solution to the reference and then repeat the whole procedure $5{,}000$ times. Figure 1 shows the probability of selecting a clustering solution with c clusters. We see that a

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

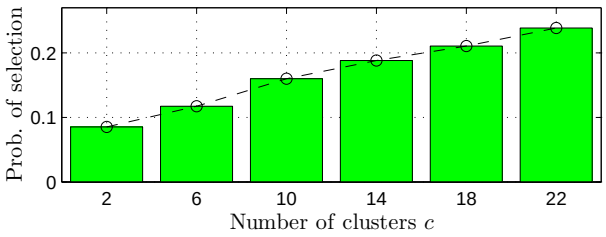

Figure 1. Probability AMI measure selects a random clustering solution with c clusters as the most similar solution to the reference clustering which has 10 clusters. AMI is biased towards selecting clustering solutions with a higher number of clusters.

clustering with 22 clusters will be selected more often than one with 2 clusters, even though we expect the former solution should be no more similar to the reference than the latter solution, due to the randomness of generation. So, although the AMI may have a constant baseline, it is still biased towards selecting clusterings containing more clusters.

To address this issue of selection bias, we go one step further in the direction of probabilistic adjustment for chance on information theoretic measures: standardizing mutual information. Using both the expected value ($E[\text{MI}]$) and the variance (Var) obtained under an assumption of random and independent clusterings, we propose the Standardized Mutual Information (SMI) measure as follows:

$$\text{SMI} = \frac{\text{MI} - E[\text{MI}]}{\sqrt{\text{Var}(\text{MI})}} \tag{1}$$

The SMI value is the number of standard deviations the mutual information is away from the mean value. This provides insight into the statistical significance of the mutual information between the clusterings. However, a key technical challenge is how to compute the variance term in the formula. We will argue that SMI has a number of desirable properties:

- When looking for the most similar clusterings to a reference one or performing external validation against a reference clustering, it reduces the bias towards selecting clusterings with more clusters or induced on fewer data points;
- The value of SMI has good interpretability, being a count of the number of standard deviations the mutual information is from the mean, under a null distribution of random clustering solutions with fixed marginals.

We advocate the use of the standardized mutual information especially when making multiple clustering comparisons in situations where the number of records is small compared to the number of clusters considered. Practically, this is the situation commonly encountered in life science and medical research. In cancer study via microarray data analysis for example, there might be up to thirteen subtypes of cancer on a data set of as few as 90 microarray samples (Monti et al., 2003). There also exist a number of important application areas such as external clustering validation, generation of alternative or multi-view clusterings (Müller et al., 2013), categorical feature selection (each feature can be seen as a clustering), and the exploration of the clustering space using results from the Meta-Clustering algorithm (Caruana et al., 2006) when the task it to find similar/dissimilar clustering from a query one.

Overall, the contributions of this paper are as follows: i) We identify new biases of the mutual information when comparing clusterings: selection bias towards clusterings with more clusters and selection bias towards clusterings induced on fewer records; ii) We propose the standardized mutual information measure (SMI) to address these bias issues; iii) To compute the SMI, we provide an analytical formula for calculating the variance of mutual information and analyze its complexity.

Acronym	Full name
MI	Mutual Information
NMI	Normalized Mutual Information
AMI	Adjusted Mutual Information
SMI	Standardized Mutual Information
VI	Variation of Information
RI	Rand Index
ARI	Adjusted Rand Index

Table 1. Acronyms used in this paper.

2. Background and Related Work

We start by reviewing the literature on related work. Next, we introduce some notation and specifically focus on information theoretic measures and literature on adjustment for chance.

2.1. Partition Comparison Measures and their Bias

Measures of agreement between two clusterings might be unfairly inflated just because of statistical fluctuations. Commonly used measures for clustering comparisons, such as the RI, show an increasing trend when the number of clusters increases even if clusterings are random and independent. In (Hubert & Arabie, 1985) it was proposed to adjust a measure M for chance as follows:

$$\frac{M - E[M]}{\max M - E[M]} \tag{2}$$

An analytical formula for the expected value is used to remove the baseline component of the measure. The expected

value for measure M is computed under the null hypothesis of random and independent clusterings and $\max M$ is an upper bound for M that acts as normalization factor. The model of randomness adopted to compute the expected value is the permutation model, also called the hypergeometric model: fixing the number of points for each cluster, partitions are generated uniformly and randomly via permutations of records. Under this assumption, the distribution of measure M is known and thus so is the expected value. It has been shown that this hypothesis behaves well in practical scenarios and it was recently employed to compute the expected value of MI in (Vinh et al., 2009).

The problem of exaggerated agreement between partitions due to chance has been also extensively studied in the decision tree literature. For each internal node in a decision tree, the best partitioning according to feature values, termed split, is selected in accordance with a splitting criterion that quantifies its predictiveness towards the target classification. Splitting criteria are in fact clustering comparison measures that aim at comparing the clustering induced by the split and the one induced by the target classification. Since the very first implementation of decision trees, ad hoc methods to reduce the bias toward selection of splits with many values have involved normalization (Quinlan, 1993). However, even with normalization, partitions with higher cardinality are more preferred (White & Liu, 1994). A key pitfall pointed out for splitting criteria is the lack of statistical significance concepts. This was formalized in (Dobra & Gehrke, 2001) where it has been proven that *"a p-value of any split criterion is a nearly unbiased criterion"*. On the other hand, the use of the p-value is controversial. It has indeed been claimed that the p-value based on the chi-square distribution for splitting criteria such as the G-statistic *"is not able to distinguish between more and less informative attributes"* (Kononenko, 1995). This behaviour is mainly due to computer precision limitations when computing the p-value for informative features to the class.

The distribution of MI could be approximated by considering the G-statistic used in goodness-of-fit tests, since the G-statistic is a scaled version of the MI. The G-statistic distribution can be approximated by a chi-square distribution, but it is well known that this approximation becomes poor when the number of objects is small in regards to the number of clusters of the clusterings compared. In particular, it is inappropriate when there exists a cluster of one clustering that shares less than 5 records with any of the clusters of the clustering compared (Agresti, 2002). Thus, its applicability for clustering comparisons is limited. Alternatively, one might attempt a brute force exact computation of the distribution of MI under the hypergeometric model of randomness, to obtain a p-value. However, this rapidly becomes infeasible, even for modest sized cases. Indeed, it is as hard as computing the p-value for the Fisher's exact test, which is not used in clustering comparison due to its computational demand. Although there exist methods to speed up Fisher's exact test using graph-based algorithms, it is still asymptotically and practically slow (Mehta & Patel, 1983). An exact p-value for MI suffers from the same problems. A common workaround is to estimate the MI distribution via Monte Carlo simulations (Frank & Witten, 1998). However, this method is still time consuming when a given degree of accuracy is required and it does not provide an exact (analytical) result for the value of the variance.

In contrast, we will shortly see that, it is possible to analytically compute the variance for mutual information under the hypergeometric model of randomness to standardize it, and this computation is orders of magnitude faster than a brute force computation of the full distribution. Moreover, a standardized measure can discriminate between clustering solutions that show high agreement with the reference clustering better than a p-value because is less prone to computer precision errors. Therefore, we use the variance and the expected value to standardize the mutual information (the SMI measure) and experimentally demonstrate that employing SMI can decrease the bias towards selecting clusterings with more clusters and towards selecting clusterings estimated on fewer data points.

2.2. Notation and Information Theoretic Measures

Let $\mathbf{A}$ and $\mathbf{B}$ be two clusterings of a dataset consisting of N records. Let $\mathbf{A}$ cluster the data in r clusters and define a_i as the size of cluster $i = 1, \ldots, r$, and let $\mathbf{B}$ cluster the data in c clusters of size b_j for each cluster $j = 1, \ldots, c$. Naturally, $\sum_{i=1}^{r} a_i = \sum_{j=1}^{c} b_j = N$. Given that $\mathbf{A}$ and $\mathbf{B}$ are partitions of the same data it is possible to count the elements that belong both to cluster i and j. Let n_{ij} denote the number of records shared between cluster i and j. The overlap between two clusterings can be represented in matrix form by a $r \times c$ contingency table $\mathcal{M}$ such as the one in Table 2. We refer to $a_i = \sum_j n_{ij}$ as the row marginals and to $b_j = \sum_i n_{ij}$ as the column marginals.

		$\mathbf{B}$				
		b_1	$\cdots$	b_j	$\cdots$	b_c
$\mathbf{A}$	a_1	n_{11}	$\cdots$	$\cdot$	$\cdots$	n_{1c}
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	a_i	$\cdot$		n_{ij}		$\cdot$
	$\vdots$	$\vdots$		$\vdots$		$\vdots$
	a_r	n_{r1}	$\cdots$	$\cdot$	$\cdots$	n_{rc}

Table 2. $r \times c$ contingency table $\mathcal{M}$ related to two clusterings $\mathbf{A}$ and $\mathbf{B}$. $a_i = \sum_j n_{ij}$ are the row marginals and $b_j = \sum_i n_{ij}$ are the column marginals.

In order to employ information theory to measure the

agreement between partitions, we have to treat the clusterings as random variables. Using the maximum likelihood estimation method, we estimate the empirical joint probability distribution of clusterings $\mathbf{A}$ and $\mathbf{B}$ as $\frac{a_i}{N}$, $\frac{b_j}{N}$, and $\frac{n_{ij}}{N}$ for the probability that an element falls in cluster i, cluster j, and both cluster i and j respectively. The entropy for a clustering is defined as the expected value of its information content if it is seen as a random variable. We can therefore define entropy for clustering $\mathbf{A}$ and $\mathbf{B}$ as follows[1]: $H(\mathbf{A}) \triangleq -\sum_{i=1}^{r} \frac{a_i}{N} \log \frac{a_i}{N}$, $H(\mathbf{B}) \triangleq -\sum_{j=1}^{c} \frac{b_j}{N} \log \frac{b_j}{N}$.

Mutual information $\mathrm{MI}(\mathbf{A}, \mathbf{B})$ quantifies the value of information shared between the two random variables and can be defined using the entropy definitions: $\mathrm{MI}(\mathbf{A}, \mathbf{B}) \triangleq H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A}, \mathbf{B}) = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j}$ where $H(\mathbf{A}, \mathbf{B})$ is the joint entropy between clusterings. Intuitively, it computes the total amount of uncertainty of each variable independently minus the uncertainty when put together. Naturally, mutual information between two clusterings can also be computed from their associated contingency table, that is $\mathrm{MI}(\mathcal{M}) = \mathrm{MI}(\mathbf{A}, \mathbf{B})$. When it is obvious, we drop the arguments and simply write MI for mutual information computed between two clusterings.

The mutual information has many possible upper bounds that might be used to obtain the Normalized Mutual Information (NMI): $\mathrm{MI}(\mathbf{A}, \mathbf{B}) \leq \min\{H(\mathbf{A}), H(\mathbf{B})\} \leq \sqrt{H(\mathbf{A}) \cdot H(\mathbf{B})} \leq \frac{1}{2}(H(\mathbf{A}) + H(\mathbf{B})) \leq \max\{H(\mathbf{A}), H(\mathbf{B})\} \leq H(\mathbf{A}, \mathbf{B})$. Depending on the chosen upper bound, it is possible to obtain information theoretic distance measures with metric properties (Vinh et al., 2010). A distance measure with metric properties is indeed useful for designing efficient algorithms that exploit the nice geometric properties of metric spaces (Meilă, 2012). An example of a true metric is the variation of information (VI), defined in (Meilă, 2007): $\mathrm{VI}(\mathbf{A}, \mathbf{B}) \triangleq H(\mathbf{A}) + H(\mathbf{B}) - 2\mathrm{MI}(\mathbf{A}, \mathbf{B})$.

In order to adjust the MI for chance as in equation (2) we have to compute the expected value over all possible contingency tables $\mathcal{M}$ with fixed number of points and fixed marginals and this is extremely time expensive. It has also been shown that the mere counting of such contingency tables with fixed marginals is $\#\mathcal{P}$-complete (Dyer et al., 1997). In (Vinh et al., 2009) the complexity of the problem has been dramatically reduced by reordering the sums in $E[\mathrm{MI}]$:

$$E[\mathrm{MI}] = \sum_{\mathcal{M}} \mathrm{MI}(\mathcal{M}) P(\mathcal{M}) = \sum_{\mathcal{M}} \sum_{i,j} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} P(\mathcal{M}) = \sum_{i,j} \sum_{n_{ij}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} P(n_{ij})$$

The inner summation varies over the support of a hypergeometric distribution and reduces the complexity to linear in the number of records, as we will show in Section 3.3. According to the adopted permutation model, N_{ij} is a hypergeometric distribution that models the sampling without replacement of a_i records among N possible ones where the number of total successes is b_j: $N_{ij} \sim \mathrm{Hyp}(a_i, b_j, N)$, $P(n_{ij}) = \frac{\binom{b_j}{n_{ij}}\binom{N-b_j}{a_i-n_{ij}}}{\binom{N}{a_i}}$, $n_{ij} \in [\max\{0, a_i + b_j - N\}, \min\{a_i, b_j\}]$. Note that if we swap a_i with b_j we obtain the same probability distribution, i.e. $\mathrm{Hyp}(a_i, b_j, N) = \mathrm{Hyp}(b_j, a_i, N)$.

AMI[2] is computed according equation (2):

$$\mathrm{AMI} = \frac{\mathrm{MI} - E[\mathrm{MI}]}{\sqrt{H(\mathbf{A}) \cdot H(\mathbf{B})} - E[\mathrm{MI}]} \tag{3}$$

3. Standardization of Information Theoretic Measures

In Section 3.1 we provide an analytical formula for the variance of mutual information under the hypergeometric model of randomness. We use the variance to standardize information theoretic measures in Section 3.2. The computational complexity of SMI is derived in Section 3.3.

3.1. Variance of Mutual Information

In order to compute the variance of MI we need to compute its second moment $E[\mathrm{MI}^2]$. For this purpose, we first set up an additional pair of indexes i' and j' in order to take care of all possible cross-products between cells:

$$E[\mathrm{MI}^2] = \sum_{\mathcal{M}} \left(\sum_{i=1}^{r} \sum_{j=1}^{c} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} \right)^2 P(\mathcal{M}) = \tag{4}$$

$$\sum_{\mathcal{M}} \sum_{i,j,i',j'} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} \cdot \frac{n_{i'j'}}{N} \log \frac{n_{i'j'} N}{a_{i'} b_{j'}} P(\mathcal{M}) =$$

$$\sum_{i,j,i',j'} \sum_{n_{ij}} \sum_{n_{i'j'}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} \cdot \frac{n_{i'j'}}{N} \log \frac{n_{i'j'} N}{a_{i'} b_{j'}} P(n_{ij}, n_{i'j'})$$

As in the mean value computation, we can swap the outer summation across all contingency tables (with fixed marginals) and sum over all possible values for cells (i, j) and (i', j'). Yet, it is difficult to compute the joint probability distribution $P(n_{ij}, n_{i'j'})$ for two general cells in the contingency table and it is necessary to treat cells differently according to their positions. Moreover, two cells belonging to two different rows and columns are inherently interacting through the remaining cells that their rows and

[1] All logarithms are considered in base 2, $\log \equiv \log_2$

[2] We choose to normalize AMI with $\sqrt{H(\mathbf{A}) \cdot H(\mathbf{B})}$ ($\mathrm{AMI}_{\mathrm{sqrt}}$) here and in the rest of the paper as proposed in (Vinh et al., 2010) given that the experimental results discussed in Section 4 are similar with other normalization factors.

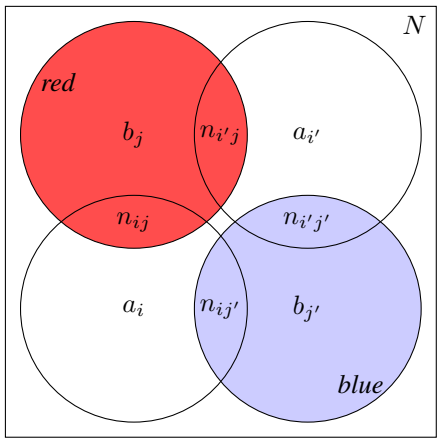

Figure 2. Urn containing N marbles among which b_j are red, and $b_{j'}$ are blue, and $N - b_j - b_{j'}$ are white.

columns have in common. This means when we consider cells (i, j) and (i', j'), we also have to take care of cells (i, j') and (i', j).

Theorem 1. *The variance of MI under the hypergeometric hypothesis is* $\mathrm{Var}(\mathrm{MI}) = E[\mathrm{MI}^2] - E[\mathrm{MI}]^2$ *where*

$$E[MI^2] = \sum_{i=1}^{r}\sum_{j=1}^{c}\sum_{n_{ij}} \frac{n_{ij}}{N}\log\frac{n_{ij}N}{a_i b_j}P(n_{ij}) \times \Bigg[$$

$$\frac{n_{ij}}{N}\log\frac{n_{ij}N}{a_i b_j} + \sum_{i'\neq i}\sum_{n_{i'j}} \frac{n_{i'j}}{N}\log\frac{n_{i'j}N}{a_{i'}b_j}P(n_{i'j}|n_{ij}) +$$

$$\sum_{j'\neq j}\sum_{n_{ij'}} P(n_{ij'}|n_{ij})\left(\frac{n_{ij'}}{N}\log\frac{n_{ij'}N}{a_i b_{j'}} + \right.$$

$$\left.\sum_{i'\neq i}\sum_{n_{i'j'}} \frac{n_{i'j'}}{N}\log\frac{n_{i'j'}N}{a_{i'}b_{j'}}P(n_{i'j'}|n_{ij'},n_{ij})\right)\Bigg]$$

$$\begin{aligned} \textit{with}\quad & N_{ij} \sim \mathit{Hyp}(a_i, b_j, N) \\ & N_{i'j}|N_{ij} \sim \mathit{Hyp}(b_j - n_{ij}, a_{i'}, N - a_i) \\ & N_{ij'}|N_{ij} \sim \mathit{Hyp}(a_i - n_{ij}, b_{j'}, N - b_j) \\ & N_{i'j'}|N_{ij'}, N_{ij} \sim \mathit{Hyp}(a_{i'}, b_{j'} - n_{ij'}, N - a_i) \end{aligned}$$

Proof. In order to make the derivation easier to follow, we employ an urn model with red, blue, and white marbles as illustrative example. Figure 2 represents such urn containing N total marbles among which b_j are red, $b_{j'}$ are blue, and $N - b_j - b_{j'}$ are white. This urn is used to simulate the sampling experiment without replacement modelled by the hypergeometric distribution. For example, it is easy to see that the random variable N_{ij} defined above models the probability of obtaining n_{ij} red marbles among a_i drawn from an urn of N marbles among which b_j are red.

We rewrite the joint probability as a product of conditional probabilities $P(n_{ij})P(n_{i'j'}|n_{ij})$. The random variable $N_{i'j'}|N_{ij}$ distributes differently depending on the possible combinations of indexes i, i', j, j':

Case 1: $i' = i \wedge j' = j$

This is the simplest case, in which $P(n_{i'j'}|n_{ij}) = 1$ if and only if $n_{i'j'} = n_{ij}$ and 0 otherwise. This case produces the first term $\frac{n_{ij}}{N}\log\frac{n_{ij}N}{a_i b_j}$ enclosed in square brackets.

Case 2: $i' = i \wedge j' \neq j$

Figure 2 comes in help when we focus on $N_{ij'}|N_{ij}$. In this case, the possible successes are the $b_{j'}$ blue marbles. We have already sampled n_{ij} red marbles and we are not interested in red marbles any more, thus the total ones available are $N - b_j$. Thus, $N_{ij'}|N_{ij} \sim \mathrm{Hyp}(a_i - n_{ij}, b_{j'}, N - b_j)$.

Case 3: $i' \neq i \wedge j' = j$

This case is symmetric to the previous one where $a_{i'}$ is now the possible number of successes. Therefore $N_{i'j}|N_{ij} \sim \mathrm{Hyp}(b_j - n_{ij}, a_{i'}, N - a_i)$.

Case 4: $i' \neq i \wedge j' \neq j$

This is the most complicated case. When all indexes are different we cannot write $N_{i'j'}|N_{ij}$ as a single hypergeometric distribution. We might think about this scenario as the second draw from the urn in Figure 2. We have already sampled a_i marbles focusing on the red ones as successes. We are now going to sample other $a_{i'}$ marbles but focusing on blue ones as successes. Just knowing that n_{ij} red ones have already been sampled does not allow us to know how many blue ones remain in the urn. Indeed, only with that information we can obtain the hypergeometric distribution. If we know that $n_{ij'}$ blue marbles have already been sampled we know there are $b_{j'} - n_{ij'}$ possible successes and thus $N_{i'j'}|N_{ij'}, N_{ij} \sim \mathrm{Hyp}(a_{i'}, b_{j'} - n_{ij'}, N - a_i)$. Finally, by the law of total probability we can obtain $P(n_{i'j'}|n_{ij}) = \sum_{n_{ij'}} P(n_{i'j'}|n_{ij'}, n_{ij})P(n_{ij'}|n_{ij})$. Note that we could have conditioned on $n_{i'j}$ and obtained the symmetric version of the above probability. The result follows from algebraic manipulations of equation (4). □

3.2. Standardized Mutual Information

We can obtain standardized versions of information theoretic measures knowing the mean value and standard deviation of the MI as per Eq. (1). An interesting point to note is that, standardization unifies several existing measures for clustering. To see this, let us define the Standardized Variation of Information (SVI) and Standardized G-statistic (SG):

$$\mathrm{SVI} = \frac{E[\mathrm{VI}] - \mathrm{VI}}{\sqrt{\mathrm{Var}(\mathrm{VI})}}, \quad \mathrm{S}G = \frac{G - E[G]}{\sqrt{\mathrm{Var}(G)}}$$

Theorem 2. *Standardization unifies the mutual information* MI*, variation of information* VI *and the* G*-statistic:* $\mathrm{SMI} = \mathrm{SVI} = \mathrm{S}G$.

Proof. $H(\mathbf{A})$ and $H(\mathbf{B})$ are constant under the fixed marginal assumption, thus the VI is a linear function of the MI under the hypergeometric hypothesis. The G-statistic is equal to a linear scaling of the MI ($G = 2N \cdot \log_e(2) \cdot \text{MI}$). The standardized version of a linear function of MI is equal to SMI because of the properties of expected value and variance. □

This 'unification' property is useful, recalling that for the normalized mutual information NMI and adjusted mutual information AMI, depending on the upperbound used, there can be as many as 5 different variants for each measure (Vinh et al., 2010).

3.3. Computational Complexity

Significant computational speedups might be obtained in computing the expected value of mutual information if probabilities are computed iteratively as: $P(n_{ij}+1) = P(n_{ij})\frac{(a_i-n_{ij})(b_j-n_{ij})}{(n_{ij}+1)(N-a_i-b_j+n_{ij}+1)}$. Here, we give a novel result characterizing the complexity of computing the expected MI:

Theorem 3. *The computational complexity for* $E[\text{MI}]$ *is* $\mathcal{O}(\max\{rN, cN\})$.

The proof is in the supplementary material. The computational complexity of computing the mean value is linear in the number of records and symmetric in c and r. This does not happen for the variance of mutual information, where we have to choose whether to condition on either $n_{ij'}$ or $n_{i'j}$ and this leads to an asymmetric computational complexity in regards to the number of rows and columns. Choosing to condition on $n_{ij'}$ as in Theorem 1, the computational complexity for SMI is dominated by the computational complexity of $E[\text{MI}^2]$:

Theorem 4. *The computational complexity for* $E[\text{MI}^2]$ *is* $\mathcal{O}(\max\{rcN^3, c^2N^3\})$.

The proof is in the supplementary material. If the number of columns c in the contingency table is greater than the number of rows r, a longer computational time is incurred. For example, if we fix the number of records N, the computation time for the variance of MI for a contingency table with $r = 6$ rows and $c = 2$ columns is bounded above by $12N^3$. Yet, for the same but transposed table (with $r = 2$ rows and $c = 6$ columns), the time is bounded by $36N^3$. Given that we can transpose a contingency table and obtain identical variance results, we can always transpose to tables where the number of rows r is higher than the number of columns c, thus making the computation faster.

4. Experiments

In this Section, we carry out several experiments to demonstrate how SMI improves interpretability as well as helping to reduce the bias toward clusterings with more clusters and towards those estimated on fewer data points.

4.1. Interpretability

We provide an example that strongly highlights the improved interpretability of SMI in comparison to the AMI. We are aiming to determine whether a clustering solution (Clustering $\mathbf{B}$) is significant compared to a random solution when performing external validation against a reference clustering (Clustering $\mathbf{A}$). Both clusterings consist of 2 clusters of equal size 50 and their cluster overlap is represented in the contingency table in Table 3. The agreement between the two clusterings is very high, indeed cluster A_1 in $\mathbf{A}$ shares 47 elements with cluster B_1 in $\mathbf{B}$, and 47 elements are also shared between cluster A_2 in $\mathbf{A}$ and B_2 in $\mathbf{B}$. Nonetheless, we obtain a modest score of 0.67 for AMI, that seems apparently far from 1, the maximum achievable. It seems that there may be plenty of clusterings $\mathbf{B}$ that could show more agreement with $\mathbf{A}$ given that 33% of the total range of values to the maximum is still possibly achievable. However, if we randomly assign records to the clusters in $\mathbf{B}$, fixing the size of each cluster to 50, we notice that it is very difficult to find a clustering whose AMI with $\mathbf{A}$ is more than 0.67. In fact, 95% of such clusterings have AMI less than 0.03. This characteristic is highlighted by SMI, which has a value of 64.22. It means that the MI between $\mathbf{A}$ and $\mathbf{B}$ is 64 standard deviations away from the mean value under the hypothesis of random and independent clusterings and therefore highly significant.

		B	
		50	50
A	50	47	3
	50	3	47

Table 3. Contingency table related to clusterings $\mathbf{A}$ and $\mathbf{B}$ that show high agreement. Nonetheless, the AMI measure is just equal to 0.67 that is apparently far from the maximum achievable 1. SMI value is 64.22 which highlights that the clustering solution $\mathbf{B}$ is significantly better than a random clustering solution.

In order to achieve even more interpretability, a p-value for mutual information might be obtained by fitting a distribution parametrized on the mean and the standard deviation. Good candidates might be the Gamma and the Normal distributions (Dobra & Gehrke, 2001; Vinh et al., 2009). However, there are no theoretical proofs about the quality of these approximations available in literature. A conservative approach we can take is to use Cantelli's inequality, which holds for all distributions: $P(\text{SMI} \geq \lambda) \leq \frac{1}{1+\lambda^2}$. This inequality states that if SMI is greater than 4.36, then the upper bound for the p-value under the hypergeometric null hypothesis is 0.05. In the above example, we get a p-value of ~ 0.0002, which again is highly significant.

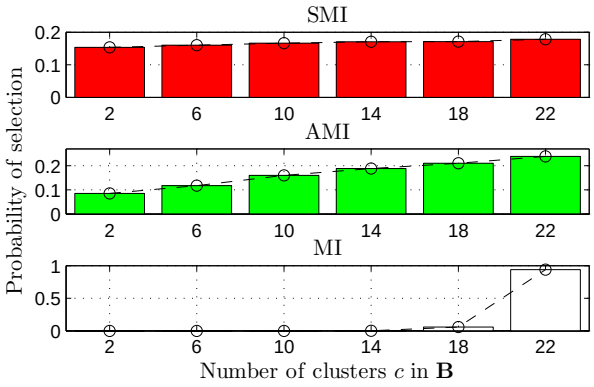

Figure 3. Estimated selection probability of a random clustering solution $\mathbf{B}$ with c clusters when compared with a reference one with 10 clusters. MI is strongly biased towards selecting clustering solutions with 22 clusters, and so is AMI, despite being baseline adjusted. SMI shows close to constant selection probability across the different solutions.

4.2. Bias Towards Clusterings with More Clusters

Consider the scenario where the user has to compare some clustering solutions to a reference one and select the one that agrees with it the most using information theoretic measures. Each solution might have been obtained using different clustering algorithms, or setting different parameters for a single algorithm of choice, e.g. varying k in k-Means. Using MI and AMI, clustering solutions with more clusters have more chances to be selected. We have observed that although the AMI has a constant baseline of 0, its variance increases as the number of clusters increases, thus creating a bias towards clustering with more clusters.

Let $\mathbf{A}$ be a reference clustering of $N = 500$ records with 10 equal size clusters. If we randomly generate clusterings $\mathbf{B}$ with different number of clusters c independently from the reference one, we do not expect any clustering solution to outperform the others in terms of agreement with $\mathbf{A}$. We carry out an experiment as follows: we generate a pool of 6 random clusterings $\mathbf{B}$ with different numbers of clusters $c = 2, 6, 10, 14, 18, 22$ and give a win to the solution that obtains the higher value for respectively SMI, AMI, and MI against the reference clustering $\mathbf{A}$. If a measure is unbiased, we expect that each clustering is selected as often as the others, that is 16.7% of the time. Figure 3 shows the estimated 'winning' frequencies obtained from 5,000 trials. We can see that random clusterings $\mathbf{B}$ with 22 clusters are selected more than 90% of the time if we use the MI. Even if we use the adjusted-for-chance AMI, such clusterings are selected 24% of the time versus the 8% for the random ones with 2 clusters. As observed, SMI helps to decrease this bias significantly. SMI shows close to constant probability of selection across different solutions but negligible differences might still exists because we are not using the full distribution of MI.

4.3. Bias Towards Clusterings Estimated on Fewer Data Points

Clustering solutions might also be induced on different numbers of data points. This is the application scenario commonly encountered in modern data processing tasks, such as streaming or distributed data. In streaming, the initial snapshots of the data often contain fewer data. Similarly, in distributed data processing, each node might have limited access to a small part of the whole data set, due to scale or privacy requirements. On the same data, one can still encounter this situation, as in the following scenario: recall that a discrete feature can be interpreted as a clustering, in which each cluster contains data points having the same feature value. Suppose we have a number of features (clusterings) and wish to compare the similarity of each against a reference clustering (class label), then choosing the feature with highest similarity to the class. If the features have missing values, then the respective clustering solutions will contain varying numbers of data points.

In these situations, there is selection bias if one uses MI and AMI as the clustering comparison measure. To demonstrate this point, we generate a random reference clustering with 4 clusters and 100 data points and then generate 5 random clustering solutions with 4 clusters, each induced using a different number of data points (20, 40, 60, 80 and 100). Each of the 5 clusterings is compared against the reference clustering (discarding from the reference any points not present in the candidate clustering solution). Even though each solution is random and independent from the reference clustering, MI and AMI select the one with 20 records significantly more often than the one with 100. Figure 4 shows the 'winning' probabilities estimated from 10,000 trials. As observed, SMI helps to decrease the bias significantly.

4.4. SMI Running Time

We compare the execution time of SMI implemented in Matlab[3] and the Fisher's exact test available in R implemented as discussed in (Mehta & Patel, 1983). We make this comparison, since Fisher's test is a very popular, yet expensive exact method and makes a good benchmark for assessing the relative runtime performance of SMI given that its computational effort is the same as for an exact p-value for MI. On a quadcore Intel Core-i7 2.9GHz PC with 16Gb of RAM, the average running time for 10 random clusterings is provided in Table 4. Each two compared clusterings were generated by assigning randomly each record to one cluster with equal probability and independently from the others. Even with a carefully-tuned

[3]The code is available at https://sites.google.com/site/icml2014smi

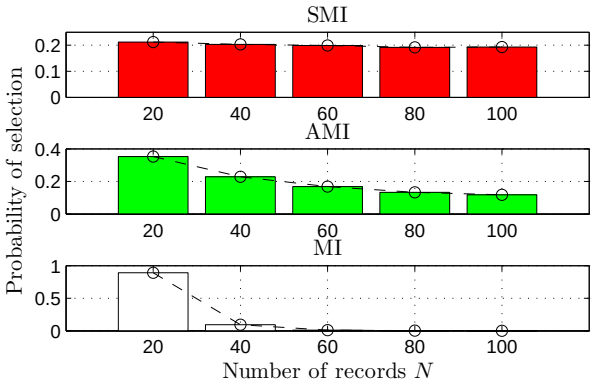

Figure 4. Estimated selection probability of a random clustering solution varying the number of points used to induce it. MI and AMI show strong bias toward selection of solutions with fewer data points. SMI shows close to constant selection probability across the solutions.

implementation, Fisher's exact test is extremely computationally expensive: it becomes intractable for a fairly small number of records N and when either the number of rows r or columns c increases. We note that computing the Fisher's exact test with the network algorithm implemented in R also requires significant memory, i.e. $\sim$ 1Gb of RAM for the data used herein.

It is worth noting that computing the SMI is highly amenable to parallelization and it is easy to implement a parallel version using modern programming languages. For example, Matlab provides the `parfor` construct that splits the load of a `for` loop on different CPUs. We can choose to parallelize the outer loops in i and j to exploits better parallelism even on 2×2 tables. It is achievable by iterating on another variable u from 1 to rc and using i and j as follows: $i \leftarrow \lceil u/c \rceil$, $j \leftarrow (u-1) \mod c + 1$. We indeed obtain almost linear speedup when $r > 2$ or $c > 2$. For example, for $r = 5$ and $c = 5$, the speedups for two and four CPU cores are 1.96 and 3.64 folds on average.

$N = 100$ records in $r \times c$ tables

	3×3	4×4	5×5	6×6	7×7	8×8
SMI	0.30	0.64	1.12	1.72	2.47	3.30
SMI $_{\text{parallel}}$	0.15	0.27	0.40	0.55	0.80	1.01
Fisher's	0.01	0.61	67.06	857.11	N/A	N/A

4×4 tables with N records

	100	150	200	250	300	350
SMI	0.65	1.53	2.94	5.00	7.59	11.00
SMI $_{\text{parallel}}$	0.30	0.51	0.97	1.52	2.33	3.35
Fisher's	0.65	11.32	242.67	844.62	N/A	N/A

Table 4. Running times in seconds for SMI and Fisher's exact test. Fisher's exact test becomes intractable when the number of records N is large or the number of rows r or columns c is large.

5. Discussion and Conclusion

In this paper, we have introduced a further degree of adjustment for chance for information theoretic measures: the standardization of mutual information. We showed that standardization unifies several well known measures, including the mutual information, the variation of information, and the G-statistic. We have provided an analytical formula to compute the variance of MI under the hypothesis of random and independent clusterings. We also analysed its computational complexity and provided running time comparisons against the Fisher's exact test as a benchmark. We experimentally demonstrated that Standardized Mutual Information (SMI) reduces the bias towards selecting clusterings with more clusters and clusterings induced on fewer data points, and arguably provides better interpretability and comparability compared to using the AMI for clustering comparison.

To conclude, we provide a radar chart to highlight the relative utility information theoretic measures in Figure 5. Each axis assesses the capability with respect to a particular clustering comparison scenario. In some situations, the user might be interested to know how far a solution is from the maximum agreement achievable with the reference clustering. In this case, VI, NMI and AMI are good choices. On the other hand, SMI is particularly useful when the task is selection of a clustering based on multiple clustering comparisons against a reference and when there are clusterings induced on data sets where the number of records is small relative to the number of clusters. As a rule of thumb, SMI should definitely be employed if $\frac{N}{r \cdot c} < 5$ and more than three clustering solutions have to be compared. Lastly, we remark this paper is focused on MI for clustering comparisons but we know that all results are applicable when using MI in arbitrary size contingency tables.

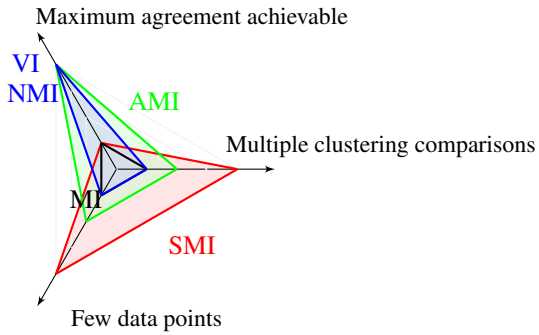

Figure 5. Relative utility of SMI, AMI, VI, NMI, MI on different clustering comparison scenario (best viewed in colors).

References

Aggarwal, C. C. and Reddy, C. K. *Data Clustering: Algorithms and Applications*. CRC Press, 2013.

Agresti, A. *Categorical data analysis*, volume 359. John Wiley & Sons, 2002.

Ben-Hur, A., Elisseeff, A., and Guyon, I. A stability based method for discovering structure in clustered data. In *Pacific symposium on biocomputing*, volume 7, pp. 6–17, 2001.

Caruana, R., Elhaway, M., Nguyen, N., and Smith, C. Meta clustering. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pp. 107–118. IEEE, 2006.

Cover, T. M. and Thomas, J. A. *Elements of information theory*. John Wiley & Sons, 2012.

Dobra, A. and Gehrke, J. Bias correction in classification tree construction. In *ICML*, pp. 90–97, 2001.

Dyer, M., Kannan, R., and Mount, J. Sampling contingency tables. *Random Structures and Algorithms*, 10(4):487–506, 1997.

Frank, E. and Witten, I. H. Using a permutation test for attribute selection in decision trees. In *ICML*, pp. 152–160, 1998.

Hubert, L. and Arabie, P. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

Kononenko, I. On biases in estimating multi-valued attributes. In *International Joint Conferences on Artificial Intelligence*, pp. 1034–1040, 1995.

Mehta, C. R. and Patel, N. R. A network algorithm for performing fisher's exact test in r $\times$ c contingency tables. *Journal of the American Statistical Association*, 78 (382):427–434, 1983.

Meilă, M. Comparing clusteringsan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007.

Meilă, M. Local equivalences of distances between clusteringsa geometric perspective. *Machine learning*, 86 (3):369–389, 2012.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1-2):91–118, 2003. ISSN 0885-6125.

Müller, E., Günnemann, S., Färber, I., and Seidl, T. Discovering multiple clustering solutions: Grouping objects in different views of the data. Tutorial at ICML, 2013. URL http://dme.rwth-aachen.de/en/DMCS.

Quinlan, J. R. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993. ISBN 1-55860-238-0.

Rand, W. M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *ICML*, pp. 1073–1080. ACM, 2009.

Vinh, N. X., Epps, J., and Bailey, J. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11:2837–2854, 2010.

White, A. P. and Liu, W. Z. Bias in information-based measures in decision tree induction. *Machine Learning*, pp. 321–329, 1994.

Wu, J., Xiong, H., and Chen, J. Adapting the right measures for k-means clustering. In *Knowledge Discovery and Data Mining*, pp. 877–886, 2009.