Tokyo Workshop on Statistically Sound Data Mining – February 16th 2015

Statistically Correcting for Chance using the Adjusted and Standardized Mutual Information Measures

James Bailey

THE UNIVERSITY OF MELBOURNE

Department of Computing and Information Systems
The University of Melbourne
Victoria, Australia

Mutual Information
- Definition
- Applications

Normalized Mutual Information
- Motivation
- Limitations

Adjusted Mutual Information
- Motivation
- Limitations

Standardized Mutual Information
- Motivation
- Characteristics of standardized measures

Conclusion
- Summary
- References

Definition of Mutual Information

Mutual Information (MI) quantifies the *information shared* between two **categorical** random variables X and Y:

$$\begin{aligned} \mathrm{MI}(X,Y) &= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x) p_Y(y)} \\ &= H(Y) - H(Y|X) \end{aligned}$$

where H is the entropy function which quantifies *uncertainty*. MI intuitively quantifies the uncertainty of Y explained by X[1].

Characteristics

- $\mathrm{MI}(X,Y) = 0$ if X and Y are independent;
- MI is maximized when one variable is a deterministic function of the other. E.g. $Y = f(X) \Rightarrow \mathrm{MI}(X,Y) = H(Y)$.

[1] In this talk we use natural logarithms.

Extension to continuous random variables

MI can also quantify the dependency between two **continuous** random variables:

$$\mathrm{MI}(X,Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}$$

Characteristics

- $\mathrm{MI}(X,Y) = 0$ if X and Y are independent;

Importance of MI

MI is a compelling tool to assess the strength of the dependency between features because it is based on a *well-established theory* and quantifies *non-linear* interactions which might be missed if e.g. the Pearson's correlation coefficient $r(X,Y)$ is used.

Estimation of MI

Categorical variables

The estimation for the categorical case is straightforward: the empirical probability distribution for $p_{X,Y}(x,y)$,$p_X(x)$, and $p_Y(y)$ is computed on data and plugged in the MI formula. In this case, MI is also a linear function of the G-statistics used in likelihood-ratio tests : $G = 2N \cdot \text{MI}$ with N number of records.

Continuous variables

A number of different estimators have been proposed for MI in the continuous case. The standard approach consists in *discretizing* the space of possible values for X and Y. There are also many possible approaches for discretization [Garcia et al., 2013], however the straightforward way is to discretize X and Y according to equal-width or equal-frequency binning.

Group	Type	Citation
Discretization based	Discretization equal width	[Steuer et al., 2002]
	Discretization equal frequency	[Steuer et al., 2002]
	Adaptive Discretization	[Cellucci et al., 2005]
Others	Nearest Neighbour	[Kraskov et al., 2004]
	Kernel Density Estimation	[Moon et al., 1995]

Table: List of possible estimators.

Non-exhaustive list of other dependency measures

Information theory gave birth to some new dependency measures (also based on discretization) in the last few years:

Acronym	Name	Citation
MIC	Maximal Information Coefficient	[Reshef et al., 2011]
GMIC	Generalized Mean Information Coefficient	[Luedtke and Tran, 2013]
MID	Mutual Information Dimension	[Sugiyama and Borgwardt, 2013]

Of course the number of possible non-linear dependency measures in use is large:

Acronym	Name	Citation
dCorr	Distance Correlation	[Székely et al., 2009]
RDC	Randomized Dependency Coefficient	[Lopez-Paz et al., 2013]
HSIC	Hilbert-Schmidt Independence Criterion	[Gretton et al., 2005]

However, information theory provides a well-established framework and it has been successfully employed for a variety of applications...

Applications

Supervised data mining

- Feature selection [Nguyen et al., 2014b, Nguyen et al., 2014a];
- Decision tree induction [Criminisi et al., 2012].

Unsupervised data mining

- External clustering validation [Romano et al., 2014];
- Generation of alternative or multi-view clusterings [Dang and Bailey, 2015, Müller et al., 2013];
- The exploration of the clustering space using results from the Meta-Clustering algorithm [Caruana et al., 2006].

Exploratory data mining

- Analysis of neural time-series data [Cohen, 2014];
- Reverse engineering of biological networks [Villaverde et al., 2013];

Application examples

Remark:

In the rest of the talk we focus on MI for **categorical** variables or the **discretized** version of continuous variables.

Examples:

To gain intuition about MI computation we describe in detail 2 application examples:

1. External clustering validation;
2. Decision tree induction.

Application example (1): external clustering validation

Task: Compare a clustering solution **B** to a reference clustering **A**.

Example

$N = 15$ data points

reference clustering **A** with 2 clusters, stars  and circles 

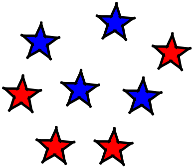
Application example (1): external clustering validation

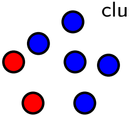
Task: Compare a clustering solution **B** to a reference clustering **A**.

Example

$N = 15$ data points

reference clustering **A** with 2 clusters, stars ☆ and circles ○

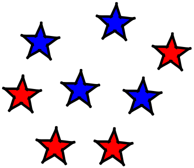


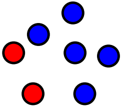


clustering solution **B** with 2 clusters, red and blue

MI computed on a contingency table

MI is estimated on data via a *contingency table* that assess the amount of overlap between **A** and **B**





		B red	**B** blue
		6	9
A ☆	8	4	4
A ○	7	2	5

MI computation

MI between the two clusterings **A** and **B** is computed on a contingency table $\mathcal{M}$ using the empirical probability distributions $\frac{n_{ij}}{N}$, $\frac{a_i}{N}$, and $\frac{b_j}{N}$:

$$\text{MI}(\mathbf{A},\mathbf{B}) = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{n_{ij}}{N} \log \frac{n_{ij}N}{a_i b_j}$$

		B				
		b_1	$\cdots$	b_j	$\cdots$	b_c
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}

Contingency table $\mathcal{M}$

$a_i = \sum_j n_{ij}$ are the row marginals and $b_j = \sum_i n_{ij}$ are the column marginals.

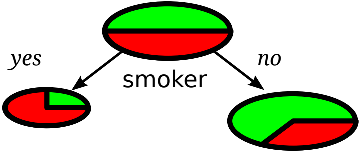
Application example (2): decision tree induction

Task: Find the most informative feature **F** to the target class **C**.

$\text{MI}(\mathbf{F}, \mathbf{C})$ is still computed on a contingency table. In this scenario MI is also known as the Information Gain: $\text{IG}(\mathbf{F}, \mathbf{C}) = \text{MI}(\mathbf{F}, \mathbf{C})$

E.g. if the class $\mathbf{C}$ = cancer and a feature $\mathbf{F}$ = smoker.

		+	-
		10	10
Smoker	8	6	2
Non smoker	12	4	8



Limitations

MI is a well-established tool to compare two random variables but it is has some limitations that can be overcome by its **statistical adjustments**.

Limitation and solution

- **Non-intuitive range of variation**

  ⇒*Solution:* the Normalized Mutual Information (NMI) [Kvalseth, 1987]; Ensure the range of the measure is in the range $[0, 1]$

- **Non-zero baseline**

  ⇒*Solution:* the Adjusted Mutual Information (AMI) [Vinh et al., 2009]; Value of measure is expected to be zero when sampling at random features to be correlated.

- **Selection bias**

  ⇒*Solution:* the Standardized Mutual Information (SMI) [Romano et al., 2014]; Avoid preferring features with many bins/categories.

Mutual Information
- Definition
- Applications

Normalized Mutual Information
- Motivation
- Limitations

Adjusted Mutual Information
- Motivation
- Limitations

Standardized Mutual Information
- Motivation
- Characteristics of standardized measures

Conclusion
- Summary
- References

Definition of the Normalized Mutual Information

Limitation of MI

MI has a non-intuitive range of variation. What does an MI of 5.6 mean ?

Solution

MI can be *normalized* by its maximum value in order to vary in the interval [0,1]:

$$\text{NMI} = \frac{\text{MI}}{\max \text{MI}}$$

Many possible upper bounds for $\text{MI}(\mathbf{A}, \mathbf{B})$:

$$\min\{H(\mathbf{A}), H(\mathbf{B})\} \leq \sqrt{H(\mathbf{A}) \cdot H(\mathbf{B})} \leq \frac{1}{2}(H(\mathbf{A})+H(\mathbf{B})) \leq \max\{H(\mathbf{A}), H(\mathbf{B})\} \leq H(\mathbf{A}, \mathbf{B})$$

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].

Normalization of Mutual Information

In [Vinh et al., 2010] we propose a review of possible normalization choices for MI.

Table: Normalization of Mutual Information.

Name	Expression	Range	Related sources
NMI_{joint}	$\frac{\text{MI}(\mathbf{A},\mathbf{B})}{H(\mathbf{A},\mathbf{B})}$	[0,1]	[Yao, 2003]
NMI_{max}	$\frac{\text{MI}(\mathbf{A},\mathbf{B})}{\max\{H(\mathbf{A}),H(\mathbf{B})\}}$	[0,1]	[Kvalseth, 1987]
NMI_{sum}	$\frac{2\text{MI}(\mathbf{A},\mathbf{B})}{H(\mathbf{A})+H(\mathbf{B})}$	[0,1]	[Kvalseth, 1987]
NMI_{sqrt}	$\frac{\text{MI}(\mathbf{A},\mathbf{B})}{\sqrt{H(\mathbf{A})H(\mathbf{B})}}$	[0,1]	[Strehl and Ghosh, 2002]
NMI_{min}	$\frac{\text{MI}(\mathbf{A},\mathbf{B})}{\min\{H(\mathbf{A}),H(\mathbf{B})\}}$	[0,1]	

Table: Distance measures based on MI.

Name	Expression	Range	Metric	Related sources
D_{joint} (VI) (Variation of Information)	$H(\mathbf{A},\mathbf{B}) - \text{MI}(\mathbf{A},\mathbf{B})$	$[0, \log N]$	✓	[Yao, 2003] [Meilă, 2005]
D_{max}	$\max\{H(\mathbf{A}),H(\mathbf{B})\} - \text{MI}(\mathbf{A},\mathbf{B})$	$[0, \log N]$	✓	
$D_{sum}(\equiv \frac{1}{2}D_{joint})$	$\frac{1}{2}[H(\mathbf{A}) + H(\mathbf{B})] - \text{MI}(\mathbf{A},\mathbf{B})$	$[0, \log N]$	✓	
D_{sqrt}	$\sqrt{H(\mathbf{A})H(\mathbf{B})} - \text{MI}(\mathbf{A},\mathbf{B})$	$[0, \log N]$	✗	
D_{min}	$\min\{H(\mathbf{A}),H(\mathbf{B})\} - \text{MI}(\mathbf{A},\mathbf{B})$	$[0, \log N]$	✗	

Successful applications and limitations

NMI has been shown to be successful in:

- Clustering comparisons scenarios [Strehl and Ghosh, 2003, Wu et al., 2009];
- Decision tree induction [Quinlan, 1993];
- Feature selection [Estévez et al., 2009].

However NMI has some limitations

NMI does not have constant **0 baseline value** for independent variables **A** and **B**.

Limitation on case study: external clustering validation

Task: Compare a clustering solution **B** to reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

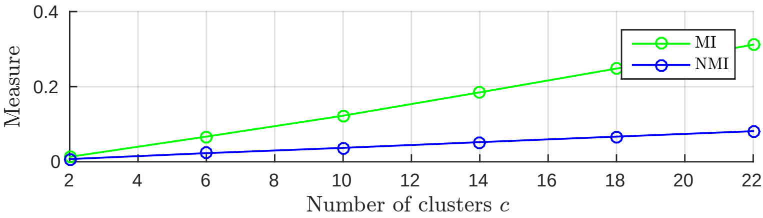


Figure: If the clustering solution **B** is generated independently from **A** at random with c clusters the average value of MI and NMI increases at the increase of the number of clusters.

Needs of statistical correction for MI

Little affect of other approaches:

A correction for MI has already been proposed a while ago [Miller, 1955]:

$$\text{MI (Miller correction)} = \text{MI} - \frac{(r-1)(c-1)}{2N}$$

with r,c number of bins and N number of records.
However it seems not effective in the general case:

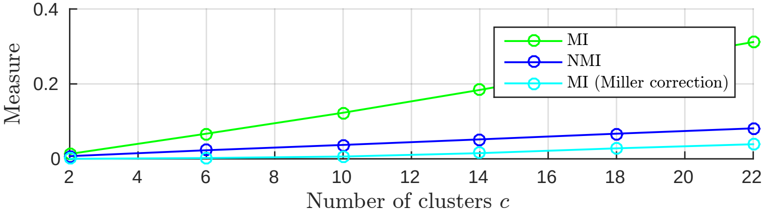


Figure: Clustering solutions **B** generated independently from **A**. Miller correction is not effective.

To address this issue we propose to statistically adjust MI for chance

Mutual Information
- Definition
- Applications

Normalized Mutual Information
- Motivation
- Limitations

Adjusted Mutual Information
- Motivation
- Limitations

Standardized Mutual Information
- Motivation
- Characteristics of standardized measures

Conclusion
- Summary
- References

The Adjusted Mutual Information

Limitation of NMI

MI and NMI have non-zero baseline.

Solution

Statistically adjust MI by the subtraction of its expected value under the null hypothesis of independence. The **Adjusted Mutual Information** (AMI) is defined as [Vinh et al., 2009]:

$$\text{AMI} = \frac{\text{MI} - E[\text{MI}]}{\max \text{MI} - E[\text{MI}]}$$

The resulting measure is statistically normalized: it is equal to 0 when MI is equal to the *expected value obtained by chance*.

Adjustment for chance

We compute the **expected value** of MI under the **null hypothesis** of independent clusterings **A** and **B**.

we make use of the **permutation model** to compute it analytically: the distribution of MI is computed using all possible contingency tables $\mathcal{M}$ obtained by permutations.

Expected Value

$E[\text{MI}]$ is obtained by summation over all possible contingency tables $\mathcal{M}$ obtained by permutations.

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

- No method to exhaustively generate $\mathcal{M}$
- extremely time expensive (permutations $\mathcal{O}(n!)$)

However, it is possible to **swap** the inner summation with the outer summation:

$$E[\text{MI}] = \underbrace{\sum_{\mathcal{M}} \sum_{i,j}}_{\textbf{to swap}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} P(\mathcal{M}) = \underbrace{\sum_{i,j} \sum_{n_{ij}}}_{\textbf{swapped}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} P(n_{ij})$$

- n_{ij} has a known **hypergeometric** distribution,
- Computation time dramatically reduced!

According to the different upper bound to MI used we obtain different versions of the Adjusted Mutual Information (AMI):

Table: Adjusted Mutual Information [Vinh et al., 2010].

Name	Expression	Range
AMI_{max}	$\frac{\text{MI}(\mathbf{A},\mathbf{B}) - E[\text{MI}(\mathbf{A},\mathbf{B})]}{\max\{H(\mathbf{A}),H(\mathbf{B})\} - E[\text{MI}(\mathbf{A},\mathbf{B})]}$	$[0, 1]^*$
AMI_{sum}	$\frac{\text{MI}(\mathbf{A},\mathbf{B}) - E[\text{MI}(\mathbf{A},\mathbf{B})]}{\frac{1}{2}(H(\mathbf{A})+H(\mathbf{B})) - E[\text{MI}(\mathbf{A},\mathbf{B})]}$	$[0, 1]^*$
AMI_{sqrt}	$\frac{\text{MI}(\mathbf{A},\mathbf{B}) - E[\text{MI}(\mathbf{A},\mathbf{B})]}{\sqrt{H(\mathbf{A})\cdot H(\mathbf{B})} - E[\text{MI}(\mathbf{A},\mathbf{B})]}$	$[0, 1]^*$
AMI_{min}	$\frac{\text{MI}(\mathbf{A},\mathbf{B}) - E[\text{MI}(\mathbf{A},\mathbf{B})]}{\min\{H(\mathbf{A}),H(\mathbf{B})\} - E[\text{MI}(\mathbf{A},\mathbf{B})]}$	$[0, 1]^*$

* These measures are normalized in a statistical sense.

Speed considerations

The computational complexity of NMI depends just on the number of clusters:

$$\mathcal{O}(rc)$$

The computational complexity of AMI is linear in the number of records N:

$$\mathcal{O}\left(\max\left\{rN, cN\right\}\right)$$

However

- Useful when the number of data points is small because

$$\lim_{N \to +\infty} E[\text{MI}] = 0$$

- Somebody has recently parallelized it [Schmidt et al., 2014].

Successful application

Task: Compare a clustering solution **B** to reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

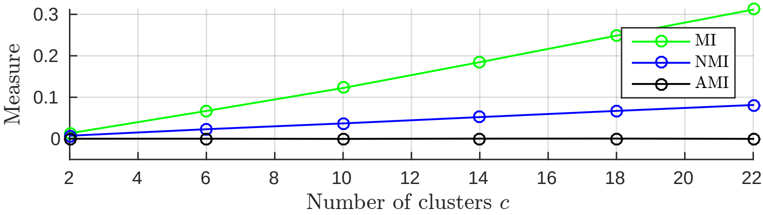


Figure: AMI obtains 0 baseline when clusterings **B** are generated at random.

Successful applications and limitations

AMI is becoming a popular tool to compare clusterings.

Title 1–20	Cited by	Year
Information theoretic measures for clusterings comparison: is a correction for chance necessary? NX Vinh, J Epps, J Bailey Proceedings of the 26th Annual International Conference on Machine Learning ...	198	2009
...		
Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance NX Vinh, J Epps, J Bailey The Journal of Machine Learning Research 11, 2837-2854	159	2010

Figure: AMI is a polar tool for clustering comparisons.

However even AMI has some limitations:
AMI is affected by **selection bias**.

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters
- One clustering solution **B** on $c = 14$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters
- One clustering solution **B** on $c = 14$ clusters
- One clustering solution **B** on $c = 18$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters
- One clustering solution **B** on $c = 14$ clusters
- One clustering solution **B** on $c = 18$ clusters
- One clustering solution **B** on $c = 22$ clusters

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters
- One clustering solution **B** on $c = 14$ clusters
- One clustering solution **B** on $c = 18$ clusters
- One clustering solution **B** on $c = 22$ clusters

Select the **B** that yields the maximum $\text{MI}(\mathbf{A}, \mathbf{B})$

Give a **win** to the solution that gets the highest value

Limitation on case study: selection of clustering solution

Task: Select the **most similar** clustering solution **B** to a reference clustering **A**.

Experiment

$N = 500$ data points
A with 10 clusters

Each **B** is generated **independently** from **A**:

- One clustering solution **B** on $c = 2$ clusters
- One clustering solution **B** on $c = 6$ clusters
- One clustering solution **B** on $c = 10$ clusters
- One clustering solution **B** on $c = 14$ clusters
- One clustering solution **B** on $c = 18$ clusters
- One clustering solution **B** on $c = 22$ clusters

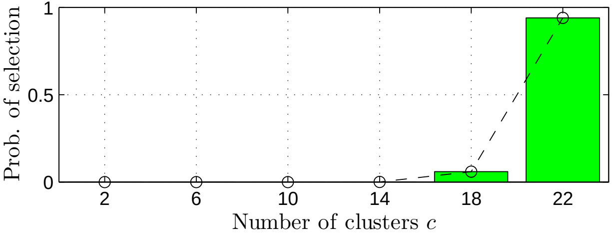
Select the **B** that yields the maximum $\mathrm{MI}(\mathbf{A}, \mathbf{B})$

Give a **win** to the solution that gets the highest value

REPEAT

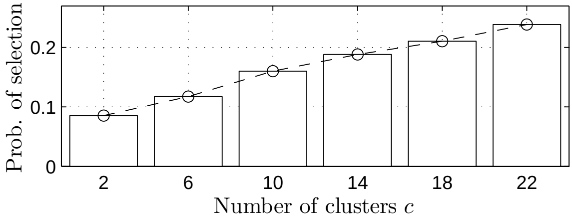
Selection Bias

MI unfairly selects more often the solution with $c = 22$ clusters.



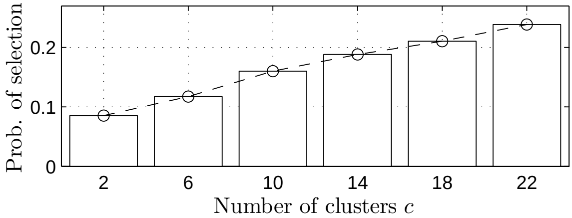
Also AMI is affected by selection bias

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



Also AMI is affected by selection bias

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



We have to take into account full distributional properties of MI: we proceed by subtracting its **expected value** and dividing by its **standard deviation**:

we propose to statistically standardize MI

Mutual Information
- Definition
- Applications

Normalized Mutual Information
- Motivation
- Limitations

Adjusted Mutual Information
- Motivation
- Limitations

Standardized Mutual Information
- Motivation
- Characteristics of standardized measures

Conclusion
- Summary
- References

Non-standardized variance

Limitation of AMI

MI, NMI, and AMI are affected by selection bias.

Solution

This behaviour is due to the non-standardized variance of AMI ⇒ need of standardization.

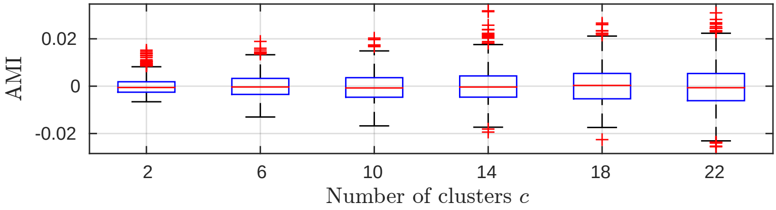


Figure: AMI values have bigger variation when the number of clusters c for **B** is high.

Definition of Standardized Mutual Information

The **Standardized Mutual Information** (SMI) is defined as [Romano et al., 2014]:

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

where we compute the **expected value** and the **variance** of Mutual Information under the **null hypothesis** of independent clusterings **A** and **B**.

The SMI value is the number of standard deviations the mutual information is away from the expected value.

As in [Vinh et al., 2009] we make use of the **permutation model** to compute the expected value and the variance:

$\Rightarrow$ The distribution of MI is computed using all possible contingency tables $\mathcal{M}$ obtained by permutations.

Variance Computation

We have to compute MI's second moment:

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

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

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

Variance Computation

We have to compute MI's second moment:

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

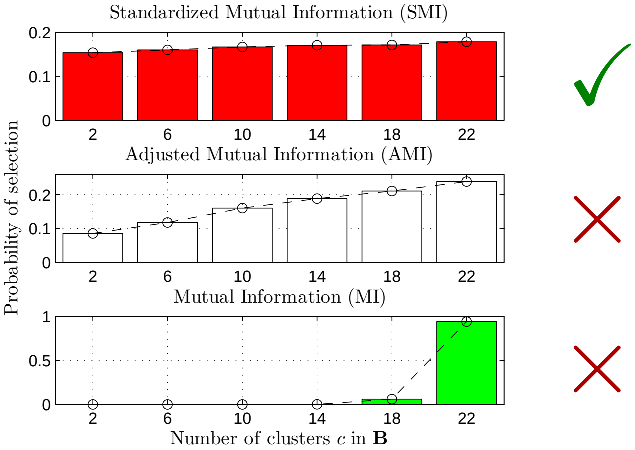
$$= \underbrace{\sum_{\mathcal{M}} \sum_{i,j,i',j'}}_{\text{to swap}} \frac{n_{ij}}{N} \log \frac{n_{ij} N}{a_i b_j} \cdot \frac{n_{i'j'}}{N} \log \frac{n_{i'j'} N}{a_{i'} b_{j'}} P(\mathcal{M})$$

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

Contribution: $P(n_{ij}, n_{i'j'})$ computation is **technically challenging**.
We use the hypergeometric model: drawings from a urn with N marbles with 3 colors, red, blue, and white.

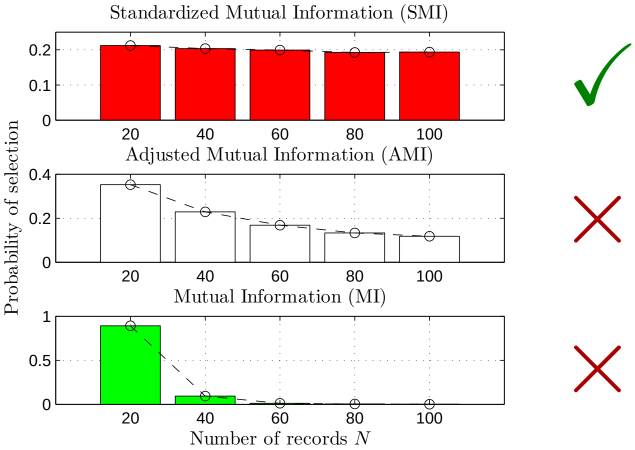
Bias Towards More Clusters Correction

MI and AMI **unfairly select** more often the solution with $c = 22$ clusters:



Bias Towards Fewer Data Points Correction

Reference clustering **A** on $N = 100$ data points with 4 clusters
B induced independently on $N = 20, 40, 60, 80, 100$ data points with 4 clusters.



Unification property

The ability to compute a variance term allows extension of the existing measures:

- Variation of Information
- G-statistic

Definitions:

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

Theorem: The standardization **unifies** information theoretic measures:

$$\text{SMI} = \text{SVI} = \text{S}G$$

Speed considerations

The computational complexity of SMI is dominated by the computational complexity of $E[\mathrm{MI}^2]$:

$$\mathcal{O}\left(\max\left\{rcN^3, c^2N^3\right\}\right)$$

However

- Useful when the number of data points is small;
- Faster than using the full distribution
  (compared to the *p*-value for the Fisher's exact test);
- Easily parallelizable.

Time in seconds for 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 (4 cores)	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

Mutual Information
Definition
Applications

Normalized Mutual Information
Motivation
Limitations

Adjusted Mutual Information
Motivation
Limitations

Standardized Mutual Information
Motivation
Characteristics of standardized measures

Conclusion
Summary
References

Summary

We discussed some enhancements to mutual information obtained by *statistical correction for chance*.

Limitation and solution

- **Non-intuitive range of variation**

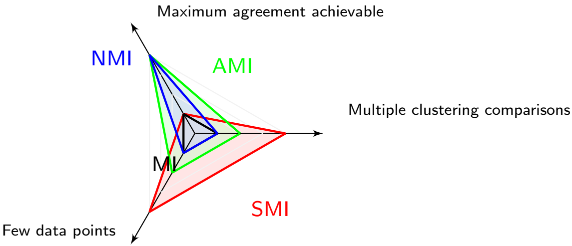
  ⇒*Solution:* the Normalized Mutual Information (NMI) [Kvalseth, 1987];
- **Non-zero baseline**

  ⇒*Solution:* the Adjusted Mutual Information (AMI) [Vinh et al., 2009];
- **Selection bias**

  ⇒*Solution:* the Standardized Mutual Information (SMI) [Romano et al., 2014];

Take Away Message

Each variant is useful in some specific scenarios and there is a trade-off in computational complexity:



Name	Range	Computational complexity
NMI	[0,1]*	$\mathcal{O}(rc)$
AMI	[0,1]	$\mathcal{O}\left(\max\{rN, cN\}\right)$
SMI	$[0,\infty)$	$\mathcal{O}\left(\max\{rcN^3, c^2N^3\}\right)$

* non statistically normalized

Table: Complexity when comparing two clusterings **A** and **B** with r and c clusters on N records.

Open issues

There is a number of open issues for SMI:

- SMI achieves strength toward selection bias at the **loss of normalization** in the range [0,1]
  $\Rightarrow$ need of statistical adjustment which allows normalization;

- SMI **computational complexity** might be problematic
  $\Rightarrow$ at the large number of records N, G-statistic ($G = 2N \cdot \text{MI}$) can be approximated with a χ^2 distribution. Need to find the scenarios where an exact SMI can be substituted by an approximation;

- SMI counts the number of standard deviations of MI, it might act as an exact p-value for MI. p-values quantifies the statistical significance of MI and this might sometimes **interfere with the effect size of MI**.

  E.g. SMI=25.4 (25.4 standard deviations away from mean). Is this closer to an *effect size* or an assessment of *statistical significance* ?
  $\Rightarrow$ need of trade-offs between importance of statistical significance and effect size.

Thank you.

Questions?

James Bailey

http://people.eng.unimelb.edu.au/baileyj/

baileyj@unimelb.edu.au

Code available online:
https://sites.google.com/site/icml2014smi/

References I

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

Cellucci, C., Albano, A. M., and Rapp, P. (2005).
Statistical validation of mutual information calculations: Comparison of alternative numerical algorithms.
Physical Review E, 71(6):066208.

Cohen, M. X. (2014).
Analyzing neural time series data: theory and practice.
MIT Press.

Criminisi, A., Shotton, J., and Konukoglu, E. (2012).
Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning.
Foundations and Trends in Computer Graphics and Vision, 7(2-3):81–227.

Dang, X. H. and Bailey, J. (2015).
A framework to uncover multiple alternative clusterings.
Machine Learning, 98(1-2):7–30.

References II

Estévez, P. A., Tesmer, M., Perez, C. A., and Zurada, J. M. (2009).
Normalized mutual information feature selection.
Neural Networks, IEEE Transactions on, 20(2):189–201.

Garcia, S., Luengo, J., Sáez, J. A., López, V., and Herrera, F. (2013).
A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning.
Knowledge and Data Engineering, IEEE Transactions on, 25(4):734–750.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005).
Measuring statistical dependence with hilbert-schmidt norms.
In *Algorithmic learning theory*, pages 63–77. Springer.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004).
Estimating mutual information.
Physical review E, 69(6):066138.

Kvalseth, T. O. (1987).
Entropy and correlation: Some comments.
Systems, Man and Cybernetics, IEEE transactions on, 17(3):517–519.

References III

Lopez-Paz, D., Hennig, P., and Schölkopf, B. (2013).
The randomized dependence coefficient.
In *Advances in Neural Information Processing Systems*, pages 1–9.

Luedtke, A. and Tran, L. (2013).
The generalized mean information coefficient.
arXiv preprint arXiv:1308.5712.

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

Meilă, M. (2005).
Comparing clusterings: an axiomatic view.
In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 577–584.

Miller, G. A. (1955).
Note on the bias of information estimates.
Information theory in psychology: Problems and methods, 2:95–100.

References IV

Moon, Y.-I., Rajagopalan, B., and Lall, U. (1995).
Estimation of mutual information using kernel density estimators.
Physical Review E, 52(3):2318.

Müller, E., Günnemann, S., Färber, I., and Seidl, T. (2013).
Discovering multiple clustering solutions: Grouping objects in different views of the data.
Tutorial at ICML.

Nguyen, X. V., Chan, J., and Bailey, J. (2014a).
Reconsidering mutual information based feature selection: A statistical significance view.
In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Nguyen, X. V., Chan, J., Romano, S., and Bailey, J. (2014b).
Effective global approaches for mutual information based feature selection.
In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 512–521. ACM.

Quinlan, J. R. (1993).
C4.5: Programs for Machine Learning.
Morgan Kaufmann.

References V

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011).
Detecting novel associations in large data sets.
Science, 334(6062):1518–1524.

Romano, S., Bailey, J., Nguyen, V., and Verspoor, K. (2014).
Standardized mutual information for clustering comparisons: One step further in adjustment for chance.
In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1143–1151.

Schmidt, T. S., Matias Rodrigues, J. F., and Mering, C. (2014).
Limits to robustness and reproducibility in the demarcation of operational taxonomic units.
Environmental microbiology.

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002).
The mutual information: detecting and evaluating dependencies between variables.
Bioinformatics, 18(suppl 2):S231–S240.

References VI

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

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

Sugiyama, M. and Borgwardt, K. M. (2013).
Measuring statistical dependence via the mutual information dimension.
In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1692–1698. AAAI Press.

Székely, G. J., Rizzo, M. L., et al. (2009).
Brownian distance covariance.
The annals of applied statistics, 3(4):1236–1265.

Villaverde, A. F., Ross, J., and Banga, J. R. (2013).
Reverse engineering cellular networks with information theoretic methods.
Cells, 2(2):306–329.

References VII

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

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

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

Yao, Y. Y. (2003).
Information-theoretic measures for knowledge discovery and data mining.
In *Entropy Measures, Maximum Entropy Principle and Emerging Applications*, pages 115–136. Karmeshu (ed.), Springer.