entropy

MDPI

Article

On Normalized Mutual Information: Measure Derivations and Properties

Tarald O. Kvålseth [1,2]

[1] Department of Mechanical Engineering, University of Minnesota, Minneapolis, MN 55455, USA; kvals001@umn.edu; Tel.: +1-952-470-1170; Fax: +1-952-470-1169

[2] Department of Industrial & Systems Engineering, University of Minnesota, Minneapolis, MN 55455, USA

Received: 26 October 2017; Accepted: 20 November 2017; Published: 22 November 2017

Abstract: Starting with a new formulation for the mutual information (MI) between a pair of events, this paper derives alternative upper bounds and extends those to the case of two discrete random variables. Normalized mutual information (NMI) measures are then obtained from those bounds, emphasizing the use of least upper bounds. Conditional NMI measures are also derived for three different events and three different random variables. Since the MI formulation for a pair of events is always nonnegative, it can properly be extended to include weighted MI and NMI measures for pairs of events or for random variables that are analogous to the well-known weighted entropy. This weighted MI is generalized to the case of continuous random variables. Such weighted measures have the advantage over previously proposed measures of always being nonnegative. A simple transformation is derived for the NMI, such that the transformed measures have the value-validity property necessary for making various appropriate comparisons between values of those measures. A numerical example is provided.

Keywords: mutual information; normalized mutual information; association measures; similarity measures; value validity

1. Introduction

Originating with the classic and profoundly influential work by Shannon [1], the *mutual information* between discrete random variables X and Y, also referred to as *transinformation* (e.g., Reza [2]), is defined as

$$I(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i, y_j)\log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) \tag{1}$$

where $p(x_i, y_j)$, $p(x_i)$, and $p(y_j)$ denote the joint and marginal probabilities and where the natural (base-e) logarithm will be used throughout this paper, although the base-2 logarithm is often used in information theory (with $\log_2 a = \log_e a/\log_e 2$). Similarly, the *conditional mutual information* between X and Y given another random variable Z is defined as

$$I(X;Y|Z) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} p(x_i, y_j, z_k)\log\left(\frac{p(x_i, y_j, z_k)p(z_k)}{p(x_i, z_k)p(y_j, z_k)}\right) \tag{2}$$

(e.g., [3] (p. 153); [4] (pp. 34–35); [5] (p. 23)).

The $I(X;Y)$ in (1) also follows from the Kullback-Leibler divergence or “statistical distance” of the probability distribution $\{p(x_i, y_j)\}$ from the corresponding independence distribution $\{p(x_i)p(y_j)\}$ [6], but this does not apply to (2). The fundamental measures in (1) and (2) also lead to various entropies and inequalities. For example, it follows from (1) that

$$\begin{aligned} I(X;Y) &= \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i, y_j)\log\left(\frac{p(x_i|y_j)}{p(x_i)}\right) = -\sum_{i=1}^{I} p(x_i)\log\, p(x_i) + \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i, y_j)\log\, p(x_i|y_j) \\ &= H(X) - H(X|Y) \end{aligned} \tag{3}$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is called the conditional entropy of X given Y. An inequality such as $H(X) \geq H(X|Y)$ follows from (3) and the fact that $I(X;Y) \geq 0$. The $I(X;Y)$ and $I(X;Y|Z)$ are often defined via entropies as in (3) rather than directly as in (1) and (2) (e.g., [7] (p. 139); [4] (p. 31)).

Since $I(X;Y)$ and $I(X;Y|Z)$ do not generally have fixed upper bounds, it is sometimes preferable to normalize those measures such that $I^*(X;Y) \in [0,\ 1]$ and $I^*(X;Y|Z) \in [0,\ 1]$. Those normalized variants, especially $I^*(X;Y)$, have been used for various purposes in a wide variety of situations such as measuring association (correlation) between X and Y (e.g., [8,9]; [10] (pp. 230–238); [11] (pp. 83–85)), similarity or performance in cluster analysis used in pattern recognition and data mining (e.g., [12,13]), non-linear dependence between X and Y using histogram-based estimation (e.g., [14,15]), and measuring performance for classifier evaluation (e.g., [16]) and of image fusion (e.g., [17]).

As a clarification of the notation used throughout this paper, $I(X;Y)$, $I^*(X;Y)$, and related symbols are used so as to be consistent with the standard notation used in information theory. Of course, neither I nor I^* are strictly functions of X or Y, but of the probability distribution $\{p(x_i, y_j)\}$ (and $p(x_i) = \sum_{j=1}^{J} p(x_i, y_j)$ and $p(y_j) = \sum_{i=1}^{I} p(x_i, y_j)$). Similarly, p is used for both the joint probability and the marginal probabilities instead of p_{XY}, p_X, and p_Y. When necessary for the sake of clarity, $I(\{p(x_i, y_j)\})$ and $I^*(\{p(x_i, y_j)\})$ are sometimes used.

As discussed in this paper, there are any number of ways of normalizing $I(X;Y)$ and $I(X;Y|Z)$, some of which result in important and unique properties. The analysis presented here is based on an alternative formulation of the mutual information between individual events $X = x_i$ and $Y = y_j$. This fundamental formulation also provides a convenient basis for introducing appropriate weighted variants of the normalized mutual information measures that, besides the probabilities, incorporate certain weights that are associated with the random variables. Furthermore, this paper discusses the important requirement that such measures need to take on numerical values that are indeed reasonable throughout the $[0,\ 1]$-interval. A simple transformation is derived to meet this requirement.

2. Mutual Information and Upper Bounds

2.1. Pairwise Measure

The $I(X;Y)$ in (1) is a weighted mean of $I(x_i;y_j) = \log[p(x_i, y_j)/p(x_i)p(y_j)]$, which is considered to be a measure of the mutual information between the two events $X = x_i$ and $Y = y_j$ or the information conveyed by $Y = y_j$ about $X = x_i$ (e.g., [2] (pp. 104–105); [3] (pp. 138–140)). This $I(x_i;y_j)$ has also been referred to as the *self-mutual information* for the event pair $(X = x_i,\ Y = y_j)$ ([4] (p. 33)).

One of the limitations of this $I(x_i;y_j)$ is that it is not necessarily nonnegative. However, an alternative nonnegative measure can be defined, starting with the well-known inequality $-\log\ c + c - 1 > 0$ for all $c > 0$ (e.g., [18] (p. 106)). Setting $c = b/a$ for $a > 0$ and $b > 0$ and multiplying each side of the inequality with a/b gives

$$f(a,b) = \frac{a}{b}\log\left(\frac{a}{b}\right) - \frac{a}{b} + 1 \geq 0 \tag{4}$$

with $f(a,b) = 0$ if, and only if, $a = b$. The function f is strictly convex in a/b since the second-order derivative $d^2 f(a,b)/d(a/b)^2 = b/a > 0$. Then, setting $a = p(x_i, y_j)$ and $b = p(x_i)p(y_j)$ in (4) results in

$$I(x_i;y_j) = \frac{p(x_i,y_j)}{p(x_i)p(y_j)} \log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) - \frac{p(x_i,y_j)}{p(x_i)p(y_j)} + 1 \geq 0 \tag{5}$$

which is proposed as a new measure of the mutual information between the two events $X = x_i$ and $Y = y_j$.

The following properties of $I(x_i;y_j)$ for all $X = x_i$ and $Y = y_j$ follow immediately from the definition in (5):

(i) $I(x_i;y_j) \geq 0$.

(ii) $I(x_i;y_j) = 0$, if, and only if, the events $X = x_i$ and $Y = y_j$ are independent.

(iii) $I(x_i;y_j) = I(y_j;x_i)$, i.e., I is symmetric in the events $X = x_i$ and $Y = y_j$.

(iv) $\sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i)p(y_j)I(x_i;y_j) = I(X;Y)$ in (1).

Note that $I(x_i;y_j)$ is also defined when $p(x_i,y_j) = 0$ (and $I(x_i;y_j) = 1$) in the limiting sense that $a \log\, a \to 0$ as $a \to 0$.

Upper bounds on $I(x_i;y_j)$ in (5) can readily be determined from the fact that $\log[p(x_i,y_j)/p(x_i)p(y_j)]$ is a strictly increasing function of $p(x_i,y_j)/p(x_i)p(y_j)$. Then, since $p(x_i,y_j) \leq p(x_i)$, it follows from (5) that

$$I(x_i;y_j) \leq U_y(x_i;y_j) = \frac{p(x_i,y_j)}{p(x_i)p(y_j)}\left[\log\left(\frac{1}{p(y_j)}\right) - 1\right] + 1 \tag{6}$$

and, since $p(x_i,y_j) \leq p(x_i)$,

$$I(x_i;y_j) \leq U_X(x_i;y_j) = \frac{p(x_i,y_j)}{p(x_i)p(y_j)}\left[\log\left(\frac{1}{p(x_i)}\right) - 1\right] + 1 \tag{7}$$

with the least upper bound being $\min\{U_X(x_i;y_j),\ U_y(x_i;y_j)\}$.

2.2. Mean Measures

The mutual information between X (or the set of events $\{X = x_i : i = 1,\ldots,I\}$) and the specific event $Y = y_j$ can logically be defined as the following weighted mean of $I(x_i;y_j)$ in (5) for $i = 1, \ldots, I$:

$$I(X;y_j) = \sum_{i=1}^{I} p(x_i)I(x_i;y_j) = \sum_{i=1}^{I} \frac{p(x_i,y_j)}{p(y_j)} \log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) \tag{8}$$

or, in terms of conditional probabilities,

$$\begin{aligned} I(X;y_j) &= -\sum_{i=1}^{I} p(x_i|y_j)\log\, p(x_i) + \sum_{i=1}^{I} p(x_i|y_j)\log\, p(x_i|y_j) \\ &= -\sum_{i=1}^{I} p(x_i|y_j)\log\, p(x_i) - H(X|y_j) \end{aligned} \tag{9}$$

where $H(X|y_j)$ is the conditional entropy of X given $Y = y_j$. The $I(x_i;Y)$ can similarly be defined as $I(x_i;Y) = \sum_{j=1}^{J} p(y_j)I(x_i;y_j)$. It follows from the properties of $I(x_i;y_j)$ that $I(X;y_j) \geq 0$ with equality when X and Y are independent.

Hamming ([3] (pp. 140–141)) uses the term conditional mutual information for $I(X;y_j)$, but this term is commonly reserved for a measure such as $I(X;Y|Z)$ in (2). From the first expression in (8), it seems that it is most appropriate to call $I(X;y_j)$ the mutual information between X and $Y = y_j$ or the information about X conveyed by the occurrence of the event $Y = y_j$.

From the expression in (8) and the upper bounds on $I(x_i;y_j)$ in (6) and (7), upper bounds on $I(X;y_j)$ are given by

$$U_y(X;y_j) = \sum_{i=1}^{I} p(x_i)U_y(x_i;y_j) = -\log\ p(y_j) \tag{10}$$

$$U_x(X;y_j) = \sum_{i=1}^{I} p(x_i)U_x(x_i;y_j) = -\sum_{i=1}^{I} p(x_i|y_j)\log\ p(x_i) \tag{11}$$

Note that the bound in (11) is equal to the first term of $I(X;y_j)$ in (9). The same bounds are also obtained by setting $p(x_i,y_j) \leq p(x_i)$ and $p(x_i,y_j) \leq p(y_j)$ in the term $\log\left[p(x_i,y_j)/p(x_i)p(y_j)\right]$ of (8).

The $I(X;Y)$ in (1) is the following weighted mean of $I(x_i;y_j)$ in (5) or of $I(X;y_j)$ in (8):

$$I(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i)p(y_j)I(x_i;y_j) = \sum_{j=1}^{J} p(y_j)I(X;y_j) \tag{12}$$

Upper bounds on $I(X;Y)$ are then obtained from (10)–(12) as

$$U_y(X;Y) = \sum_{j=1}^{J} p(y_j)U_y(X;y_j) = -\sum_{j=1}^{J} p(y_j)\log\ p(y_j) = H(Y) \tag{13}$$

$$U_x(X;Y) = \sum_{j=1}^{J} p(y_j)U_x(X;y_j) = -\sum_{i=1}^{I} p(x_i)\log\ p(x_i) = H(X) \tag{14}$$

The same bounds are also obtained by setting $p(x_i,y_j) \leq p(x_i)$ and $p(x_i,y_j) \leq p(y_j)$ in the term $\log\left[p(x_i,y_j)/p(x_i)p(y_j)\right]$ in (1).

2.3. Conditional Measures

In the case of three random variables X, Y, and Z with conditional probabilities $p(x_i|z_k)$, $p(y_j|z_k)$, and $p(x_i,y_j|z_k)$ with $k = 1,\ldots,K$, one can define the mutual information between the events $X = x_i$ and $Y = y_j$ conditional on the event $Z = z_k$ by setting $a = p(x_i,y_j|z_k)$ and $b = p(x_i|z_k)p(y_j|z_k)$ in (4) so that

$$I(x_i;y_j|z_k) = \frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}\left[\log\left(\frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}\right) - 1\right] + 1 \tag{15}$$

where $I(x_i;y_j|z_k) \geq 0$ with equality only under conditional independence, i.e., if, and only if, $p(x_i;y_j|z_k) = p(x_i|z_k)p(y_j|z_k)$.

The mutual information between X and Y given $Z = z_k$ can then be defined as the following weighted mean of $I(x_i;y_j|z_k)$

$$\begin{aligned} I(X;Y|z_k) &= \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i|z_k)p(y_j|z_k)I(x_i;y_j|z_k) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i,y_j|z_k)\log\left(\frac{p(x_i;y_j|z_k)}{p(x_i|z_k)\,p(y_j|z_k)}\right) \end{aligned} \tag{16}$$

The conditional mutual information of X and Y given Z as defined in (2) follows from (16) as

$$I(X;Y|Z) = \sum_{k=1}^{K} p(z_k)I(X;Y|z_k) \tag{17}$$

Since $\log(a/b)$ is a strictly increasing function of a/b, the following upper bounds on $I(x_i;y_j|z_k)$ in (15) are obtained from $p(x_i,y_j|z_k) \leq p(x_i|z_k)$ and $p(x_i,y_j|z_k) \leq p(x_i|z_k)$:

$$I(x_i;y_j|z_k) \leq U_y(x_i;y_j|z_k) = \frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}\left[\log\left(\frac{1}{p(y_j|z_k)}\right)-1\right]+1 \tag{18}$$

$$I(x_i;y_j|z_k) \leq U_x(x_i;y_j|z_k) = \frac{p(x_i,y_j|z_k)}{p(x_i|z_k)p(y_j|z_k)}\left[\log\left(\frac{1}{p(x_i|z_k)}\right)-1\right]+1 \tag{19}$$

From (16), (18), and (19), upper bounds on $I(X;Y|z_k)$ are given by

$$\begin{aligned} U_y(X;Y|z_k) &= \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i|z_k)\,p(y_j|z_k)U_y(x_i;y_j|z_k) \\ &= -\sum_{j=1}^{J} p(y_j|z_k)\log p(y_j|z_k) = H(Y|z_k) \end{aligned} \tag{20}$$

$$\begin{aligned} U_x(X;Y|z_k) &= \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i|z_k)p(y_j|z_k)U_x(x_i;y_j|z_k) \\ &= -\sum_{i=1}^{I} p(x_i|z_k)\log p(x_i|z_k) = H(X|z_k) \end{aligned} \tag{21}$$

From (17), (20), and (21), upper bounds on $I(X;Y|Z)$ become

$$U_y(X;Y|Z) = \sum_{k=1}^{K} p(z_k)U_y(X;Y|z_k) = H(Y|Z) \tag{22}$$

$$U_x(X;Y|Z) = \sum_{k=1}^{K} p(z_k)U_x(X;Y|z_k) = H(X|Z) \tag{23}$$

3. Normalizations

Let I denote any one of the mutual information measures in (5), (8), (12), and (15)–(17) with its derived upper bounds U_x and U_y and let

$$I^* = I/U \in [0,\ 1] \tag{24}$$

denote a normalized form of I. Either $U = U_x$ or $U = U_y$ would satisfy (24). However, there exists literally infinitely many potential candidates for U in (24), as represented by the α-order arithmetic mean

$$U_\alpha = \left(\frac{U_x^\alpha + U_y^\alpha}{2}\right)^{1/\alpha},\ \alpha \in (-\infty,\infty) \tag{25}$$

where α is some real-valued parameter. For any given (fixed) U_x and U_y, U_α is a nondecreasing function of α and is strictly increasing unless $U_x = U_y$ ([19] (pp. 16–18)). Other means could also be considered, such as the logarithmic mean and Stolarsky means [20,21]. See also [16].

Particularly well-known members of U_α are the following:

$$\begin{aligned} U_{-\infty} = \lim_{\alpha\to-\infty} U_\alpha = \min\{U_x,U_y\},\ U_{-1} = \tfrac{2U_xU_y}{U_x+U_y},\ U_0 = \lim_{\alpha\to 0} U_\alpha = \sqrt{U_xU_y} \\ U_1 = \tfrac{U_x+U_y}{2},\ U_2 = \left(\tfrac{U_x^2+U_y^2}{2}\right)^{1/2},\ U_\infty = \lim_{\alpha\to\infty} U_\alpha = \max\{U_x,U_y\} \end{aligned} \tag{26}$$

in increasing order of magnitude unless $U_x = U_y$ (when they are all equal). All I^* from (24)–(26) are symmetric in X and Y. In its strong favor, $I^* = I/U_{-\infty}$ is the only member of (24) and (25) that is always capable of attaining the maximal value of 1.

In the case of $I(X;Y)$ with $U_x = H(X)$ and $U_y = H(Y)$ in (13) and (14), the most apparent normalized candidates are perhaps the following [9]:

$$I^*_{(1)}(X;Y) = \frac{I(X;Y)}{\min\{H(X), H(Y)\}},\ I^*_{(2)}(X;Y) = \frac{2I(X;Y)}{H(X)+H(Y)}$$

$$I^*_{(3)}(X;Y) = \frac{I(X;Y)}{\max\{H(X), H(Y)\}} \tag{27}$$

Horibe [8] proved that $1 - I^*_{(3)}(X;Y)$ is a distance metric so that $I^*_{(3)}(X;Y)$ becomes a (normalized) similarity metric [22]. The $I^*_{(2)}(X;Y)$ gives equal weight to H(X) and H(Y), as does $I(X;Y)/\sqrt{H(X)H(Y)}$ (see also U_{-1} and U_2 in (26)). This $I(X;Y)/\sqrt{H(X)H(Y)}$, which has been suggested by Strehl and Ghosh [23], is somewhat analogous to the correlation coefficient $\rho = \mathrm{Cov}(X;Y)/\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}$. As stated above with respect to $I/U_{-\infty}$ from (24) and (26), the $I^*_{(1)}(X;Y)$ in (27) is the single normalized $I(X;Y)$ that is always capable of attaining the value of 1.

As a further explanation of the last statement, consider the condition that either (a) for each i (i = 1, . . . , I), $p(x_i, y_j) > 0$ for at most one j or (b) for each j $(j = 1, \ldots, J)$, $p(x_i, y_j) > 0$ for at most one i. In terms of a contingency table with row variable X and column variable Y so that $p(x_i, y_j)$ is the probability in row i and column j, this condition means that either (a) each row or (b) each column contains at most one nonzero $p(x_i, y_j)$. The term "at most" is not needed if all of the marginal probabilities $p(x_i)$ and $p(y_j)$ are nonzero. For any given marginal distributions $\{p(x_i)\}$ and $\{p(y_j)\}$, this condition is clearly the one for which the mutual information (dependence, association) is at its maximum. No other $\{p(x_i, y_j)\}$ distribution could plausibly or intuitively produce a larger $I(X;Y)$. Under this condition, irrespective of the values of I and J (dimensions of the $I \times J$ contingency table), $I(X;Y) = \min\{H(X), H(Y)\}$ so that $I^*_{(1)}(X;Y) = 1$, whereas all of the other normalized variants of $I(X;Y)$, including $I^*_{(2)}(X;Y)$ and $I^*_{(3)}(X;Y)$ in (27), take on values that are necessarily less than 1. Assuming that all of the marginal probabilities are nonzero, it is only when I = J (square contingency tables) that those other normalized variants of $I(X;Y)$ are able to attain their upper bound of 1.

Consequently, unless there are particular or compelling reasons to the contrary, normalizations of mutual information measures ought to be based on the least upper bounds. Thus, for the general formulation in (24), $U = \min\{U_x, U_y\}$ should be the standard normalizing factor so that

$$I^* = \frac{I}{\min\{U_x, U_y\}} \in [0,\ 1] \tag{28}$$

This will ensure that the attainable maximal value of I^* is 1, irrespective of the marginal probabilities and the dimensions I and J.

4. Weighted Mutual Information

The idea of weighted entropy introduced by Belis and Guiasu [24] and Guiasu ([25] (Chapter 4)) and extended to include weighted divergence ([26]; [27] (pp. 33–91)) has also been formulated for mutual information as

$$I_w(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} w(x_i, y_j)p(x_i, y_j)\log\left(\frac{p(x_i, y_j)}{p(x_i)p(y_j)}\right) \tag{29}$$

where $w(x_i, y_j)$ are some nonnegative weights that are associated with the random variables X and Y [28–30]. Of course, when $w(x_i, y_j) = 1$ for all i and j, (29) reduces to (1).

A limitation of this $I_w(X;Y)$ and perhaps one reason for its relatively limited use is the fact that $I_w(X;Y)$ is not necessarily nonnegative [30,31]. One way to overcome this limitation is to restrict the weighting function w, such that it depends only on one of the two variables X and Y [30]. However, such a restriction is not necessary if the weighted mutual information measures are based on the nonnegative $I(x_i;y_j)$ in (5), such that

$$I_w(x_i;y_j) = w(x_i;y_j)I(x_i;y_j) \geq 0,\ i = 1,\ldots,I,\ j = 1,\ldots,J \tag{30}$$

and from which

$$\begin{gathered} I_w(X;y_j) = \sum_{i=1}^{I} p(x_i)I_w(x_i;y_j) \\ = \sum_{i=1}^{I} w(x_i,y_j)\frac{p(x_i,y_j)}{p(y_j)}\left[\log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) - 1\right] + \sum_{i=1}^{I} w(x_i,y_j)p(x_i) \end{gathered} \tag{31}$$

and

$$\begin{gathered} I_w(X;Y) = \sum_{j=1}^{J} p(y_j)I_w(X;y_j) \\ = \sum_{i=1}^{I}\sum_{j=1}^{J} w(x_i,y_j)p(x_i,y_j)\left[\log\left(\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\right) - 1\right] + \sum_{i=1}^{I}\sum_{j=1}^{J} w(x_i,y_j)p(x_i)p(y_j) \end{gathered} \tag{32}$$

all of which are nonnegative. When $w(x_i;y_j) = 1$ for all i and j, (30)–(32) reduce to (5), (8), and (12), respectively.

In the case of conditional mutual information measures, nonnegative weighted equivalents can be derived by starting with

$$I_w(x_i;y_j|z_k) = w(x_i,y_j,z_k)I(x_i;y_j|z_k) \tag{33}$$

for $i = 1,\ldots,I$, $j = 1,\ldots,J$, $k = 1,\ldots,K$, and for $I(x_i;y_j|z_k)$ in (15). The equivalents for (16) and (17) are then obtained from (33) as

$$I_w(X;Y|z_k) = \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i|z_k)p(y_j|z_k)I_w(x_i;y_j|z_k) \tag{34}$$

and

$$I_w(X;Y|Z) = \sum_{k=1}^{K} p(z_k)I_w(X;Y|z_k) \tag{35}$$

These weighted conditional measures are nonnegative since $I(x_i;y_j|z_k) \geq 0$ and it is assumed that the weights $w(x_i,y_j,z_k) \geq 0$ for $i = 1,\ldots,I$, $j = 1,\ldots,J$, and $k = 1,\ldots,K$.

Upper bounds and normalizations for the weighted measures in (30)–(35) can be derived in the same way as done above for their unweighted equivalents. Consider, for example, the $I_w(X;Y)$ in (32). Since the log( ) is a strictly increasing function and since $p(x_i,y_j) \leq p(x_i)$, it follows from (32) that

$$I_w(X;Y) \leq U_{wy}(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} w(x_i,y_j)p(x_i,y_j)\left[-\log\ p(y_j) - 1 + \frac{p(x_i)p(y_j)}{p(x_i,y_j)}\right] \tag{36}$$

and, since $p(x_i,y_j) \leq p(y_i)$

$$I_w(X;Y) \leq U_{wx}(X;Y) = \sum_{i=1}^{I}\sum_{j=1}^{J} w(x_i,y_j)p(x_i,y_j)\left[-\log\ p(x_i) - 1 + \frac{p(x_i)p(y_j)}{p(x_i,y_j)}\right] \tag{37}$$

For $w(x_i, y_j) = 1$ for all i and j, these upper bounds reduce to those in (13) and (14). The normalized form of $I_w(X;Y)$ with 1 as the attainable maximum value is given by

$$I_w^*(X;Y) = \frac{I_w(X;Y)}{\min\{U_{wx}(X;Y), U_{wy}(X;Y)\}} \in [0,\ 1] \tag{38}$$

with the denominator terms defined by (36) and (37). If all $w(x_i, y_j)$ are the same, not necessarily 1, (38) becomes the $I^*_{(1)}(X;Y)$ in (27).

In the case when X and Y are continuous random variables, equivalent mutual information measures to all of those introduced above for the discrete case can be obtained by substituting probability density functions f for all of the probabilities $p(\ )$, and by substituting definite integrals for the summations. Thus, the equivalent of $I_w(x_i, y_j)$ in (30) and (5) becomes

$$I_w(x;y) = w(x,y)\frac{f(x,y)}{f(x)f(y)}\left[\log\left(\frac{f(x,y)}{f(x)f(y)}\right) - 1\right] + w(x,y) \geq 0 \tag{39}$$

and the equivalent of $I_w(X;Y)$ in (32) becomes

$$\begin{aligned} I_w(X;Y) &= \int_x \int_y f(x)f(y)I_w(x;y)dxdy \\ &= \int_x \int_y w(x,y)f(x,y)\log\left(\tfrac{f(x,y)}{f(x)f(y)}\right)dxdy \\ &+ \int_x \int_y w(x,y)[f(x)f(y) - f(x,y)]dxdy \end{aligned} \tag{40}$$

where the integrals are over the entire range of values of X and Y. When $w(x,y) = 1$ for all x and y, the nonnegative $I_w(X;Y)$ in (40) reduces to the well-known mutual information

$$I(X;Y) = \int_x \int_y f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy \tag{41}$$

which is also nonnegative since $I_w(X;Y)$ is nonnegative for all $w(x,y) \geq 0$.

However, mutual information measures in the continuous case, such as those in (39)–(41) cannot generally be normalized to the [0, 1]-interval unless particular constraints are imposed. Such continuous measures do not generally have fixed upper bounds. If, for example, X and Y have the joint normal distribution with correlation coefficient ρ, then $I(X;Y) = -\log\sqrt{1-\rho^2}$ (e.g., [2] (pp. 282–283)), which increases without an upper bound as $\rho \to 1$.

5. Value Validity

5.1. Value-Validity Consideration

Let I^* stand for any one of the normalized mutual information measures discussed above, and let i_a^*, i_b^*, i_c^*, etc. stand for its numerical values for different probability distributions. It is then generally of interest to make different types of comparisons between such numerical values. While there may be no particular reason to doubt the validity of size (order) comparisons such as $i_a^* > i_b^*$, more specific comparisons such as the difference comparisons $i_a^* - i_b^* > i_c^* - i_d^*$ or $i_a^* - i_b^* = k(i_c^* - i_d^*)$ (for constant k) may require certain restrictions or modifications on I^* in order to be valid. The same type of validity requirement would apply to the interpretations of absolute values of I^*.

Although there are different types of *validity* that are used in measurement theory ([32] (Chapter 4)), *value validity* of a measure is used here to mean that all of the potential values of the measure provide true or realistic representations of the extent of the attribute being measured as supported by a generally acceptable criterion or condition. Such an analysis has been done for the normalized entropy $H^*(X) = -\sum_{i=1}^{I} p(x_i) \log\ p(x_i)/\log\ I$ [33], but a different approach is needed for the mutual information between two or more random variables.

One approach is to consider the binary random variables $X = x_1, x_2$ and $Y = y_1, y_2$ with all marginal probabilities equal to 1/2 and with the following joint probability distribution $\{p_{ij}^{\alpha}\}$:

$$\begin{array}{l} p_{11}^{\alpha} = \frac{1+\alpha}{4},\ p_{12}^{\alpha} = \frac{1-\alpha}{4} \\ p_{21}^{\alpha} = \frac{1-\alpha}{4},\ p_{22}^{\alpha} = \frac{1+\alpha}{4} \end{array} \tag{42}$$

where $\alpha \in [0,\ 1]$ is a real-valued parameter. Furthermore, consider the normalized mutual information measure I^* in (28) that takes on values between 0 and 1, inclusive. Then, for any joint distribution $\{p(x_i, y_j)\}$ with $i = 1, \ldots, I$ and $j = 1, \ldots, J$, the following equality exists:

$$I^*(\{p(x_i, y_j)\}) = I^*\left(\{p_{ij}^{\alpha}\}\right) = g(\alpha) \tag{43}$$

where g is a single-valued function of α. As a consequence of (43), the value validity of I^* for any $\{p(x_i, y_j)\}$ can be considered based on $\{p_{ij}^{\alpha}\}$.

The I^* takes on its extremal values when $\alpha = 0$ and $\alpha = 1$ with

$$I^*\left(\{p_{ij}^0\}\right) = 0,\ I^*\left(\{p_{ij}^1\}\right) = 1 \tag{44}$$

where $\{p_{ij}^0\}$ corresponds to the statistical independence condition and $\{p_{ij}^1\}$ corresponds to the complete dependence condition ($p_{11}^1 = p_{22}^1 = 1/2$ and $p_{12}^1 = p_{21}^1 = 0$). The probability distribution $\{p_{ij}^{\alpha}\} = (p_{11}^{\alpha}, p_{12}^{\alpha}, p_{21}^{\alpha}, p_{22}^{\alpha})$ can be considered as a point (or vector) in four-dimensional Euclidean space with Cartesian coordinates $p_{12}^{\alpha}, \ldots, p_{22}^{\alpha}$. Then, first, the $\{p_{ij}^{\alpha}\}$ in (42) is seen to be the weighted mean of $\{p_{ij}^0\}$ and $\{p_{ij}^1\}$ as follows:

$$\{p_{ij}^{\alpha}\} = \alpha\{p_{ij}^1\} + (1-\alpha)\{p_{ij}^0\},\ \alpha \in [0,\ 1] \tag{45}$$

Second, with I_v^* denoting a normalized mutual information measure that has the value-validity property and in terms of the Euclidean distance $d(\ )$, the following equality between distance ratios is propounded as a logical relationship:

$$\frac{\left|I_v^*\left(\{p_{ij}^{\alpha}\}\right) - I_v^*\left(\{p_{ij}^0\}\right)\right|}{\left|I_v^*\left(\{p_{ij}^1\}\right) - I_v^*\left(\{p_{ij}^0\}\right)\right|} = \frac{d\left(\{p_{ij}^{\alpha}\}, \{p_{ij}^0\}\right)}{d\left(\{p_{ij}^1\}, \{p_{ij}^0\}\right)} \tag{46}$$

Since $d\left(\{p_{ij}^{\alpha}\}, \{p_{ij}^0\}\right) = \{2[(1+\alpha)/4 - 1/4]^2 + 2[(1-\alpha)/4 - 1/4]^2\}^{1/2} = \alpha/2$ and $d\left(\{p_{ij}^1\}, \{p_{ij}^0\}\right) = 1/2$, (46) can be expressed as

$$I_v^*\left(\{p_{ij}^{\alpha}\}\right) = \alpha I_v^*\left(\{p_{ij}^1\}\right) + (1-\alpha) I_v^*\left(\{p_{ij}^0\}\right) \tag{47}$$

and, with $I_v^*\left(\{p_{ij}^0\}\right) = 0$ and $I_v^*\left(\{p_{ij}^1\}\right) = 1$ as in (44) and (47) reduces to

$$I_v^*\left(\{p_{ij}^{\alpha}\}\right) = \alpha \tag{48}$$

The value-validity condition in (47) and (48) is also a logical implication from (45). That is, $I_v^*\left(\{p_{ij}^{\alpha}\}\right)$ in (47) as a weighted mean of $I_v^*\left(\{p_{ij}^1\}\right)$ and $I_v^*\left(\{p_{ij}^0\}\right)$ is equivalent to the weighted mean of the probabilities in (45).

In the case when $\alpha = 1/2$, $p_{ij}^{1/2} = \left(p_{ij}^0 + p_{ij}^1\right)/2$ and $\left|p_{ij}^{1/2} - p_{ij}^0\right| = \left|p_{ij}^{1/2} - p_{ij}^1\right|$ for $i,\ j = 1,\ 2$ so that the distance $d\left(\{p_{ij}^{1/2}\}, \{p_{ij}^0\}\right) = d\left(\{p_{ij}^{1/2}\}, \{p_{ij}^1\}\right)$, $I_v^*\left(\{p_{ij}^{1/2}\}\right) = 1/2$ from (48), which is clearly the only logical value for a measure that can vary from 0 for $\{p_{ij}^0\}$ to 1 for $\{p_{ij}^1\}$. However, as discussed next for the normalized mutual information measures I^* in (28), $I^*\left(\{p_{ij}^{1/2}\}\right) << 1/2$ and $I^*\left(\{p_{ij}^{\alpha}\}\right)$ does not meet the value-validity condition in (48) without some required correction.

5.2. Value-Validity Corrections of I^*

Values of the normalized mutual information measures collectively included in I^* in (28) and specifically defined in (5)–(14) have been computed for the joint probability distribution $\{p_{ij}^{\alpha}\}$ in (42) and for different values of α. For each pair of upper bounds U_x and U_y in (6)–(14), $U_x = U_y$ since all of the marginal probabilities for the distribution in (42) equal 1/2 The results are summarized in Table 1.

Table 1. Values of the normalized forms of the measures in (1), (5), and (8) for the probability distribution $\{p_{ij}^{\alpha}\}$ in (42) with differing α-values.

I^*	α				
	0.1	**0.3**	**0.5**	**0.7**	**0.9**
$I^*(x_1;y_1) = I^*(x_2;y_2)$	0.01	0.07	0.20	0.42	0.77
$I^*(x_1;y_2) = I^*(x_2;y_1)$	0.01	0.06	0.18	0.37	0.69
$I^*(X;y_1) = I^*(X;y_2)$	0.01	0.07	0.19	0.39	0.71
$I^*(X;Y)$	0.01	0.07	0.19	0.39	0.71

It is clear from these results that all of the I^* measures fail to comply with the value-validity condition in (48). Their values are substantially smaller than the α-values, implying that those measures substantially understate the true extent of the normalized mutual information attribute or characteristic. The absolute extent of this understatement is greatest around the true midrange ($\alpha \approx 0.5$), while the relative understatement is greatest at the lower end (smaller α-values).

Rather than rejecting these I^* measures because of their lack of value validity and hence their restricted utility, they can be corrected or modified so as to comply with the requirement in (48) by the use of the relationship in (43). Thus, with I^* denoting any one of the measures in Table 1 and for any given joint probability distribution $\{p(x_i, y_j)\}$ for $i = 1, \ldots, I$ and $j = 1, \ldots, J$, the value of α can be determined so as to comply with the equality in (43). The solution $\alpha = I_C^*(\{p(x_i, y_j)\})$ then becomes the corrected value of I^*. Formally stated,

$$\alpha = I_C^*(\{p(x_i, y_j)\}) = h[I^*(\{p(x_i, y_j)\})], h = g^{-1} \tag{49}$$

where h is the inverse function of g in (43). This corrected I_C^* will necessarily comply with the value-validity condition in (48).

For any given distribution $\{p(x_i, y_j)\}$, the corrected value $I_C^*(\{p(x_i, y_j)\})$ of $I^*(\{p(x_i, y_j)\})$ can be obtained by using a computer search algorithm to find the value of α for which $I^*(\{p(x_i, y_j)\}) = I^*\left(\{p_{ij}^{\alpha}\}\right)$ for the $\{p_{ij}^{\alpha}\}$ in (42). The resulting α-value, which can be determined to any degree of accuracy, is then the corrected $I_C^*(\{p(x_i, y_j)\})$. Alternatively, the function g in (43) and hence h in (49) may be determined analytically, such that

$$h\left[I^*\left(\{p_{ij}^{\alpha}\}\right)\right] = \alpha \tag{50}$$

It is desirable that the function h be relatively simple and convenient to use rather than being a complex expression that is derived from some model or curve-fitting program.

Consider the data in Table 2 for $I^*(X;Y)$ based on the distribution $\{p_{ij}^{\alpha}\}$ in (42) and different values of α. By exploring various forms of h in (50), Table 2 presents the results for two potential candidates for h as approximations to the equality in (50). The square-root function mentioned in [34] does provide quite respectable approximations, i.e., $\sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)} \approx \alpha$. In fact, for the fitted model $\hat{\alpha} = \sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)}$ and the data in Table 2, the coefficient of determination, when properly computed [35], is found to be $R^2 = 1 - \sum\left(\alpha - \sqrt{I^*}\right)^2 / \sum(\alpha - \overline{\alpha})^2 = 0.97$.

Table 2. Values of I^* for $I^*(X;Y) = I(X;Y)/\min\{H(X), H(Y)\}$ and the distribution $\{p_{ij}^{\alpha}\}$ in (42) with differing α-values, as well as the corresponding values for two different functions h satisfying (50), approximately.

$\boldsymbol{\alpha}$	$\boldsymbol{I^*\left(\{p_{ij}^{\alpha}\}\right)}$	$\boldsymbol{\sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)}}$	$\boldsymbol{1 - \left(1 - \sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)}\right)^{11/9}}$
0	0	0	0
0.1	0.0072	0.0849	0.1027
0.2	0.0291	0.1706	0.2044
0.3	0.0659	0.2567	0.3041
0.4	0.1187	0.3445	0.4033
0.5	0.1887	0.4344	0.5017
0.6	0.2781	0.5274	0.5999
0.7	0.3902	0.6247	0.6981
0.8	0.5310	0.7287	0.7970
0.9	0.7136	0.8447	0.8974
1	1	1	1

However, a superior approximation can be achieved by means of regression analysis. Thus, for the simple model $(1 - \alpha) = \left(1 - \sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)}\right)^{\beta}$ and for the data in Table 2, estimated $\beta = 1.2315$, or, when rounded off to the nearest fraction, $\beta = 11/9$. The resulting function h in (50), i.e., $1 - \left(1 - \sqrt{I^*\left(\{p_{ij}^{\alpha}\}\right)}\right)^{11/9}$ is seen from the results in Table 2 to satisfy (50) to a high degree of approximation. In fact, if the data are rounded off to the second decimal place, which is clearly sufficient for most practical purposes, it is seen from Table 2 that the equality in (50) holds exactly.

Consequently, it follows from (49) that the corrected value of I^* becomes

$$I_C^*(\{p(x_i, y_j)\}) = 1 - \left(1 - \sqrt{I^*(\{p(x_i, y_j)\})}\right)^{11/9} \tag{51}$$

which complies with the value-validity condition in (48) to a high degree of approximation. Although the final part of the analysis has been based specifically on the normalized form of $I(X;Y)$ in (1), the value-validity correction in (51) is also applicable to the other normalized mutual information measures, such as $I^*(x_i; y_j)$ and $I^*(X; y_j)$, as discussed above and subject to the normalization in (28). This proposition is supported by the fact that, as indicated by the data in Table 1, all of the normalized measures deviate from the value-validity condition in (48) to a comparable extent.

5.3. Numerical Example

Table 3 gives the real sample results of United States Senate elections (for four different years) based on data given by Reynolds ([36] (p. 2)). Here $X = x_1$ is the event that a vote is for a Democratic candidate and $X = x_2$ that it is for the Republican candidate. The variable Y refers to the three parties with which the voters were identified. Based on the sample probability distribution $\{p(x_i, y_j)\}$ in Table 3, the values of the various normalized mutual information measures have been computed, as presented in Table 3. The normalizations have all been based on the least upper bounds as in (28).

The values of both the (uncorrected) measures I^* and the value-validity corrected measures I_C^* from (51) are given in Table 3.

Table 3. United States (U.S.) Senate election results in terms of sample probabilities (proportions) $p(x_i, y_j)$ for candidate vote (X) and voters' party identification (Y) (sample size N = 2843). Source: Reynolds ([36] (p. 2)).

Vote (X)	**Party Identification (Y)**			
	Democrat (y_1)	**Independent (y_2)**	**Republican (y_3)**	**Total**
Democrat (x_1)	0.39	0.11	0.04	0.54
Republican (x_2)	0.07	0.12	0.27	0.46
Total	0.46	0.23	0.31	1.00

Corresponding values for the normalized mutual information measures defined in the text:
$I^*(x_1; y_j)$ = 0.35, 0.01, 0.46; $I_C^*(x_1; y_j)$ = 0.66, 0.12, 0.75 for j = 1, 2, 3
$I^*(x_2; y_j)$ = 0.33, 0.01, 0.55; $I_C^*(x_2; y_j)$ = 0.65, 0.13, 0.81 for j = 1, 2, 3
$I^*(X; y_j)$ = 0.33, 0.01, 0.49; $I_C^*(X; y_j)$ = 0.65, 0.13, 0.77 for j = 1, 2, 3
$I^*(X; Y)$ = 0.31; $I_C^*(X; Y)$ = 0.63

The information measures may in this case be considered as measures of association (dependence, correlation) between the two categorical variables X and Y. Thus, from the overall measure $I_C^*(X; Y) = 0.63$, one can justifiably make the interpretation that there is a "somewhat high" or "substantial" degree of association between the party identification or affiliation of voters and of candidates. Or, in information-theory terminology, the (amount of) information about the vote (X) obtained by knowing the voters' party identification is "somewhat high". A similar numerical result is obtained from Cramér's coefficient of association V (e.g., [36] (p. 47)), with $V = 0.62$ for the $\{p(x_i, y_j)\}$ distribution in Table 3. However, a very different and misleading result and interpretation would be obtained if based on the $I^*(X; Y) = 0.31$ in Table 3.

A more detailed explanation about the association between X and Y can be gleaned from the $I_C^*(x_i, y_j)$ and $I_C^*(X; y_j)$ in Table 3. Both of the values $I_C^*(x_i, y_2)$ for i = 1, 2 and $I_C^*(X; y_2)$ show that relatively little information about the vote (X) is obtained from knowing that a (randomly selected) voter was an Independent ($Y = y_2$). Due to the value-validity property of I_C^* and from the results that $I_C^*(x_i; y_j) \geq 5 I_C^*(x_i, y_2)$ for $i = 1, 2$ and $j = 1, 3$, and $I_C^*(X; y_j) \geq 5 I_C^*(X; y_2)$ for $j = 1, 3$, it is permissible to infer that at least five times as much information about the vote (X) is obtained by knowing that a voter was a Democrat than if the voter was an Independent. The same inference applies to the voter being a Republican versus an Independent. As another observation, the largest pairwise $I_C^*(x_i; y_j)$ corresponds to the events $X = x_2$ (vote was for Republican candidate) and $Y = y_3$ (voter was a Republican) with $I_C^*(x_2; y_3) = 0.81$. That is, the event $Y = y_3$ provides a "very large" amount of information about the event $X = x_2$. This is significantly (about 23%) more than the $I_C^*(x_1; y_1) = 0.66$ (indicating a somewhat greater party loyalty by Republicans).

The $I_C^*(X; y_j)$ and $I_C^*(x_i; y_j)$, especially perhaps $I_C^*(X; y_j)$, are likely to be particularly useful when Y is an explanatory variable and X is a response variable. In this example, with $I_C^*(X; y_1) = 0.65$ and $I_C^*(X; y_3) = 0.77$, it can be concluded that the information about the vote (X) gained by knowing that a (randomly chosen) voter was a Republican was somewhat larger (by nearly 20%) than by knowing that the voter was a Democrat. Of course, for a small 2×3 table, as in Table 3, some of the above observations or results are rather apparent in general terms from the probabilities in Table 3, but the use of I_C^* provides a means of quantifying those observations (results).

6. Conclusions

For the potential normalizations of various mutual information measures that are discussed in this paper, the least upper bounds have been emphasized as for I^* in (28). This provides I^* with the desirable property that its upper limit of 1 can always be attained for any marginal probability distributions of the random variables X and Y (and Z in the conditional case), and for any dimensions I and J (and K in the conditional case). Such a property is generally required of any measure of association for categorical variables (e.g., [37] (Chapter 33)). In the case of $I^*(X;Y)$ (i.e., $I^*_{(1)}(X;Y)$ in (27)), another normalized form could be considered, such as $I(X;Y)/\min\{\log I, \log J\}$ [15], but this measure can only attain the value 1 when the smallest of the entropies $H(X)$ and $H(Y)$ involves equal (uniform) marginal probabilities.

It has also been emphasized above that, for comparisons other than size (order) comparisons such as $I^*(X_1;Y_1) > I^*(X_2;Y_2)$ for pairs of random variables (X_1, Y_1) and (X_2, Y_2), it is required that a measure have the value-validity property. Otherwise, results and conclusions may be incorrect and misleading. A simple transformation or correction of I^* into I^*_C provides for such a requirement. This more informative measure I^*_C permits its numerical values to be properly interpreted as to their absolute magnitudes and to be compared, so as to truly represent the attribute (characteristic) being measured.

Besides the fact that it is preferable and more convenient to interpret and compare results that vary over a fixed interval such as [0, 1], a clear advantage of using I^*_C (or I^*) over I is that I^*_C (or I^*) controls or adjusts for the size of a data set. This makes it possible to compare the results for data sets of varying size (dimension). Such control (adjustment) can be achieved directly by using the normalizing denominator $\min\{\log I, \log J\}$, or indirectly, as argued in this paper, via the marginal probability distributions of the random variables.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, 623–656. [CrossRef]
2. Reza, F.M. *An Introduction to Information Theory*; McGraw-Hill: New York, NY, USA, 1961.
3. Hamming, R.W. *Coding and Information Theory*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1980.
4. Han, T.S.; Kobayashi, K. *Mathematics of Information and Coding*; American Mathematical Society: Providence, RI, USA, 2002.
5. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; Wiley: New York, NY, USA, 2006.
6. Kullback, S.; Leibler, R.A. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [CrossRef]
7. MacKay, D.J.C. *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
8. Horibe, Y. Entropy and correlation. *IEEE Trans. Syst. Man Cybern.* **1985**, *SMC-15*, 641–642. [CrossRef]
9. Kvålseth, T.O. Entropy and correlation: Some comments. *IEEE Trans. Syst. Man Cybern.* **1987**, *SMC-17*, 517–519.
10. Wickens, T.D. *Multiway Contingency Tables Analysis for the Social Sciences*; Lawrence Erlbaum: Hillsdale, NJ, USA, 1989.
11. Tang, W.; He, H.; Tu, X.M. *Applied Categorical and Count Data Analysis*; CRC Press: Boca Raton, FL, USA, 2012.
12. Pfitzer, D.; Leibbrandt, R.; Powers, D. Characterization and evaluation of similarity measures of pairs of clusterings. *Knowl. Inf. Syst.* **2009**, *19*, 361–394. [CrossRef]
13. Yang, Y.; Ma, Z.; Yang, Y.; Nie, F.; Shen, H.T. Multitask spectral clustering by exploring intertask correlation. *IEEE Trans. Cybern.* **2015**, *45*, 1069–1080. [CrossRef] [PubMed]
14. Jain, N.; Murthy, C.A. A new estimate of mutual information based measure of dependence between two variables: Properties and fast implementation. *Int. J. Mach. Learn. Cybern.* **2015**, *7*, 857–875. [CrossRef]
15. Reshef, D.N.; Reshef, Y.A.; Finucane, H.K.; Grossman, S.R.; McVean, G.; Turnbaugh, P.J.; Lander, E.S.; Mitzenmacher, M.; Sabeti, P.C. Detecting novel associations in large data sets. *Science* **2011**, *334*, 1518–1524. [CrossRef] [PubMed]
16. Hu, B.-G. Information measure toolbox for classifier evaluation on open source software scilab. In Proceedings of the 2009 IEEE International Workshop on Open-Source Software for Scientific Computing, (OSSC-2009), Guiyang, China, 18–20 September 2009; pp. 179–184.

17. Hossny, M.; Nahavandi, S.; Creighton, D. Comments on 'Information measure for performance of image fusion'. *Electron. Lett.* **2009**, *44*, 1066–1067. [CrossRef]
18. Hardy, G.H.; Littlewood, J.E.; Pólya, G. *Inequalities*; Cambridge University Press: Cambridge, UK, 1934.
19. Beckenbach, E.F.; Bellman, R. *Inequalities*; Springer: Heidelberg, Germany, 1971.
20. Stolarsky, K.B. Generalizations of the logarithmic mean. *Math. Mag.* **1975**, *48*, 87–92. [CrossRef]
21. Ebanks, B. Looking for a few good means. *Am. Math. Mon.* **2012**, *119*, 658–669. [CrossRef]
22. Chen, S.; Ma, B.; Zhang, K. On the similarity metric and the distance metric. *Theor. Comput. Sci.* **2009**, *410*, 2365–2376. [CrossRef]
23. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
24. Belis, M.; Guiasu, S. A quantitative-qualitative measure in cybernetic systems. *IEEE Trans. Inf. Theory* **1968**, *14*, 593–594. [CrossRef]
25. Guiasu, S. *Information Theory with Applications*; McGraw-Hill: New York, NY, USA, 1977.
26. Taneja, H.C.; Tuteja, R.K. Characterization of a quantitative-qualitative measure of relative information. *Inf. Sci.* **1984**, *33*, 217–222. [CrossRef]
27. Kapur, J.N. *Measures of Information and Their Applications*; Wiley Eastern: New Delhi, India, 1994.
28. Luan, H.; Qi, F.; Xue, Z.; Chen, L.; Shen, D. Multimodality image registration by maximization of quantitative-qualitative measures of mutual information. *Pattern Recognit.* **2008**, *41*, 285–298. [CrossRef]
29. Schaffernicht, E.; Gross, H.-M. Weighted mutual information for feature selection. In Proceedings of the 21st International Conference on Artificial Neural Networks, Part II, Espoo, Finland, 14–17 June 2011; pp. 181–188.
30. Pocock, A.C. Feature Selection via Joint Likelihood. Ph.D. Thesis, School of Computer Science, University of Manchester, Manchester, UK, 2012.
31. Kvålseth, T.O. The relative useful information measure: Some comments. *Inf. Sci.* **1991**, *56*, 35–38. [CrossRef]
32. Hand, D.J. *Measurement Theory and Applications*; Wiley: Chichester, UK, 2004.
33. Kvålseth, T.O. Entropy evaluation based on value validity. *Entropy* **2014**, *16*, 4855–4873. [CrossRef]
34. Kvålseth, T.O. Association measures for nominal categorical variables. In *International Encyclopedia of Statistical Science*; Lovric, M., Ed.; Springer: Berlin/Heidelberg, Germany, 2011; Part 1, pp. 61–64.
35. Kvålseth, T.O. Cautionary note about R^2. *Am. Stat.* **1985**, *39*, 279–285.
36. Reynolds, H.T. *The Analysis of Cross-Classification*; The Free Press: New York, NY, USA, 1977.
37. Kendall, M.; Stuart, A. *The Advanced Theory of Statistics, Volume 2: Inference and Relationships*, 4th ed.; Charles Griffin: London, UK, 1979.

© 2017 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).