A General Measure of Rule Interestingness

Szymon Jaroszewicz, Dan A. Simovici

January 15, 2002

Abstract

The paper presents a new general measure of rule interestingness. Many known measures such as χ^2, gini gain or entropy gain can be obtained from this measure by setting some numerical parameters representing the amount of trust we have in the estimates of certain probabilities from the data. Moreover we show that there is a continuum of measures having χ^2, Gini gain and entropy gain as boundary cases. Properties and experimental evaluation of the new measure are also presented.

Keywords: interestingness measure, distribution, Cziser divergence Kullback-Leibler divergence, rule.

1 Introduction

Determining the interestingness of rules is an important data mining problem. Many data mining algorithms produce enormous amounts of rules, making it impossible for the user to analyze all of them by hand. It is thus essential to establish some measure by which rules interestingness can be expressed numerically and used, for example, to sort the discovered rules.

Many such measures have been proposed, and used in literature (see [1] for a survey). In this paper we concentrate on measures that assess how much knowledge we gain on the joint distribution of a set of attributes Q from the knowing the joint distribution of some set of attributes P.

Examples of such measures are *entropy gain*, *mutual information*, *Gini gain*, χ^2 [7, 9, 3, 1, 11, 10]. The rules considered here are thus different from *association rules* studied in data mining, since we consider full joint distributions of both antecedent and consequent, while association rules consider only the probability of all attributes having some specified value. This approach has the advantage of natural applicability to mulitvalued attributes.

In this paper we demonstrate that all the above mentioned measures are special cases of a more general parametric measure of interestingness, and by choosing two numerical parameters a continuum of measures can be obtained containing several well-known interesting measures as special cases.

Next, we give some essential definitions.

Definition 1 *A* probability distribution *is a matrix of the form*

$$\Delta = \begin{pmatrix} x_1 & \cdots & x_m \\ p_1 & \cdots & p_m \end{pmatrix},$$

where $p_i \geq 0$ *for* $1 \leq i \leq m$ *and* $\sum_{i=1}^{m} p_i = 1$.

Δ *is an* uniform distribution *if* $p_1 = \cdots = p_m = \frac{1}{m}$. *An* m*-valued uniform distribution will be denoted by* $\mathcal{U}_m$.

Let $\tau = (T, H, \rho)$ be a database table, where T is the name of the table, H is its heading, and ρ is its content. If $A \in H$ is an attribute of τ, the domain of A in τ is denoted by $\mathrm{dom}(A)$. The projection of a tuple $t \in \rho$ on a set of attributes $L \subseteq H$ is denoted by $t[L]$. For more on relational notation and terminology see [13].

Definition 2 *The* distribution of a set of attributes $L = \{A_1, \ldots, A_n\}$ *is the matrix*

$$\Delta_{L,\tau} = \begin{pmatrix} \ell_1 & \cdots & \ell_r \\ p_1 & \cdots & p_r \end{pmatrix}, \tag{1}$$

where $r = \prod_{j=1}^{n} |\mathrm{dom}(A_j)|$, $\ell_i \in \mathrm{dom}(A_1) \times \cdots \times \mathrm{dom}(A_n)$, *and* $p_i = \frac{|t \in \rho | t[L] = \ell_i|}{|\rho|}$ *for* $1 \leq i \leq r$.

The subscript τ will be omitted when the table τ is clear from context.

Suppose that the distribution of the attribute set L in the table $\tau = (T, H, \rho)$ is

$$\Delta_L = \begin{pmatrix} \ell_1 & \cdots & \ell_r \\ p_1 & \cdots & p_r \end{pmatrix}.$$

The *Havrda-Charvát* α*-entropy* of the attribute set L (see [6]) is defined as:

$$\mathcal{H}_\alpha(L) = \frac{1}{1-\alpha}\left(\sum_{j=1}^{r} p_j^\alpha - 1\right).$$

The limit case, when α tends towards 1 yields the Shannon entropy:

$$\mathcal{H}(L) = -\sum_{j=1}^{r} p_j \log p_j$$

Another important case is obtained when $\alpha = 2$. In this case, we obtain the Gini index of L (see [1]) given by:

$$\mathrm{gini}(L) = 1 - \sum_{j=1}^{r} p_j^2.$$

If L, K are two sets of attributes of a table τ that have the distributions

$$\Delta_L = \begin{pmatrix} l_1 & \cdots & l_m \\ p_1 & \cdots & p_m \end{pmatrix}, \text{ and } \Delta_K = \begin{pmatrix} k_1 & \cdots & k_n \\ q_1 & \cdots & q_n \end{pmatrix},$$

then the conditional Shannon entropy of L conditioned upon K is given by

$$\mathcal{H}(L|K) = -\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_j},$$

where $p_{ij} = \frac{|\{t \in \rho | t[L]=\ell_i \text{ and } t[K]=k_j\}|}{|\rho|}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$. Similarly, the Gini conditional index of these distributions is:

$$\text{gini}(L|K) = 1 - \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_{ij}^2}{q_j}.$$

These definitions allow us to introduce the Shannon gain (called entropy gain in literature [7]) and the Gini gain defined as:

$$\begin{aligned} \mathsf{gain}_{\text{gini}}(L,K) &= \text{gini}(L) - \text{gini}(L|K), \\ \mathsf{gain}_{\text{shannon}}(L,K) &= \mathcal{H}(L) - \mathcal{H}(L|K) \\ &= \mathcal{H}(L) + \mathcal{H}(K) - \mathcal{H}(L \cup K), \end{aligned} \tag{2}$$

respectively.

Notice that the Shannon gain is identical to the *mutual information* between attribute sets P and Q [7]. For the Gini gain we can write:

$$\mathsf{gain}_{\text{gini}}(L,K) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_{ij}^2}{q_j} - \sum_{i=1}^{m} p_i^2 \tag{3}$$

The product of the distributions Δ_P, Δ_Q, where

$$\Delta_P = \begin{pmatrix} x_1 & \cdots & x_m \\ p_1 & \cdots & p_m \end{pmatrix}, \text{ and } \Delta_Q = \begin{pmatrix} y_1 & \cdots & y_n \\ q_1 & \cdots & q_n \end{pmatrix},$$

is the distribution

$$\Delta_P \times \Delta_Q = \begin{pmatrix} (x_1,y_1) & \cdots & (x_m,y_n) \\ p_1 q_1 & \cdots & p_m q_n \end{pmatrix}.$$

The attribute sets P, Q are *independent* if $\Delta_{PQ} = \Delta_P \times \Delta_Q$, where PQ is an abbreviation for $P \cup Q$.

Definition 3 *A* rule *is a pair of attribute sets (P,Q). If $P, Q \subseteq H$, where $\tau = (T, H, \rho)$ is a table, then we refer to (P,Q) as a rule of τ.*

If (P,Q) is a rule, then we refer to P as the antecedent *and to Q as the* consequent *of the rule. A rule (P,Q) will be denoted, following the prevalent convention in the literature, by $P \to Q$.*

This broader definition of rules originates in [3], where rules were replaced by dependencies in order to capture statistical dependence in both the presence and absence of items in itemsets. The significance of this dependence was measured by the χ^2 test, and our approach is a further extension of that point of view.

The notion of *distribution divergence* is central to the rest of the paper.

Definition 4 *Let $\mathcal{D}$ be the class of distributions. A* distribution divergence *is a function $D : \mathcal{D} \times \mathcal{D} \longrightarrow \mathbb{R}$ such that:*

1. *$D(\Delta, \Delta') \geq 0$ and $D(\Delta, \Delta') = 0$ if and only if $\Delta = \Delta'$ for every $\Delta, \Delta' \in \mathcal{D}$.*
2. *When Δ' is fixed, $D(\Delta, \Delta')$ is a convex function of Δ; in other words, if $\Delta = a_1\Delta_1 + \cdots + a_k\Delta_k$, where $a_1 + \ldots + a_k = 1$, then*

$$D(\Delta, \Delta') \geq \sum_{i=1}^{k} a_i D(\Delta_i, \Delta').$$

An important class of distribution divergences was obtained by Cziszar in [4] as:

$$D_\phi(\Delta, \Delta') = \sum_{i=1}^{n} q_i \phi\left(\frac{p_i}{q_i}\right),$$

where

$$\Delta = \begin{pmatrix} k_1 & \cdots & k_n \\ p_1 & \cdots & p_n \end{pmatrix}, \text{ and } \Delta' = \begin{pmatrix} l_1 & \cdots & l_n \\ q_1 & \cdots & q_n \end{pmatrix},$$

are two distributions and $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ is a twice differentiable convex function such that $\phi(1) = 0$. We will also make an additional assumption that $0 \cdot \phi(\frac{0}{0}) = 0$ to handle the case when for some i both p_i and q_i are zero. If for some i, $p_i > 0$, and $q_i = 0$ the value of $D_\phi(\Delta, \Delta')$ is undefined.

The Cziszar divergence satisfies properties (1) and (2) given above (see [6]).

The following result shows the invariance of Cziszar divergence with respect to distribution product:

Theorem 1 *For any distributions Γ, Δ, Δ' and any Cziszar divergence measure D_ϕ we have $D_\phi(\Gamma \times \Delta, \Gamma \times \Delta') = D_\phi(\Delta, \Delta')$.*

Proof. Let

$$\Delta = \begin{pmatrix} k_1 & \cdots & k_n \\ p_1 & \cdots & p_n \end{pmatrix}, \Delta' = \begin{pmatrix} l_1 & \cdots & l_n \\ q_1 & \cdots & q_n \end{pmatrix}, \text{ and } \Gamma = \begin{pmatrix} h_1 & \cdots & h_m \\ r_1 & \cdots & r_m \end{pmatrix}.$$

The definition of Cziszar divergence implies

$$\begin{aligned} D_\phi(\Gamma \times \Delta, \Gamma \times \Delta') &= \sum_{i=1}^{n}\sum_{j=1}^{m} q_i r_j \phi\left(\frac{p_i r_j}{q_i r_j}\right), \\ &= \sum_{j=1}^{m} r_j \sum_{i=1}^{n} q_i \phi\left(\frac{p_i}{q_i}\right), \\ &= D_\phi(\Delta, \Delta'), \end{aligned}$$

which is the desired equality. (Q.E.D.)

Depending on the choice of the function ϕ we obtain the divergences shown in the table below:

$\phi(x)$	$D(\Delta, \Delta')$	Divergence
$x \log x$	$p_i \log \frac{p_i}{q_i}$	Kullback-Leibler
$x^2 - x$	$\sum_{i=1}^{n} \frac{p_i^2}{q_i} - 1$	D_{χ^2}

Both the Kullback-Leibler divergence (also known as *crossentropy*), which we will denote by D_{KL} and the χ^2-divergence denoted by D_{χ^2} are special cases of the Havrda-Charvát divergence $D_{\mathcal{H}_\alpha}$ generated by $\phi(x) = \frac{x^\alpha - x}{\alpha - 1}$ [6]; specifically, D_{χ^2} is obtained by taking $\alpha = 2$, while D_{KL} is obtained as a limit case, when α tends towards 1.

It is easy to verify that

$$D_{\chi^2}(\Delta, \Delta') = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}.$$

Note that $|\rho| D_{\chi^2}$ equals the χ^2 dependency measure, well known from statistics [1].

2 Interestingness of Rules

The main goal of this paper is to present a unified approach to the notion of interestingness of rules.

Let $r = P \to Q$ be a rule in a table $\tau = (T, H, \rho)$.

To construct an interestingness measure we will use a Bayesian approach, in that we will consider an assumed apriori distribution Θ of the consequent set of attributes Q. This could be an *apriori* distribution of Q, the distribution Δ_Q observed from the data, or some combination of these distributions. To define an interestingness measure of r we will be guided by two main considerations:

- The more the observed joint distribution of PQ diverges from the product distribution of P and the assumed distribution Θ of Q the more interesting the rule is. Note that $\Delta_{PQ} = \Delta_P \times \Theta$ corresponds to the situation when P and Q are independent and the observed distribution of Q follows the assumed one.
- The rule is not interesting if P, Q are independent. Therefore, we need to consider a correcting term in the definition of an interestingness measure that will decrease its value when Δ_Q is different from the assumed distribution.

The choice of the distribution Θ of the consequent Q of rules of the form $P \to Q$ can be made starting either from the content of the table, that is, adopting Δ_Q for Θ, or from some exterior information. For example, if Q is the sex attribute for a table that contains data concerning some experiment subjects, we can adopt as the assumed distribution either

$$\Delta_{sex} = \begin{pmatrix} \text{'F'} & \text{'M'} \\ 0.45 & 0.55 \end{pmatrix},$$

assuming that 45% of the individuals involved are female, or the distribution

$$\Delta_{gen\_pop} = \begin{pmatrix} \text{'F'} & \text{'M'} \\ 0.51 & 0.49 \end{pmatrix},$$

consistent with the general distribution of the sexes in the general population. Moreover, we can contemplate a convex combination of distribution of the form

$$\Theta_a = a\Delta_Q + (1-a)\Theta_0,$$

where Δ_Q is the distribution of Q that is extracted from a table τ and Θ_0 is a distribution that is based on some prior knowledge. The number a reflects the degree of confidence in the data contained by the table τ; the closer this number is to 1, the higher the confidence, and the more preponderant Δ_Q is in the assumed distribution.

Definition 5 *Let $r : P \to Q$ be a rule, D be some measure of divergence between distributions, and let Θ be a distribution.*

The measure of interestingness *generated by D and Θ is defined by*

$$\Upsilon_{D,\Theta}(r) = D(\Delta_{PQ}, \Delta_P \times \Theta) - D(\Delta_Q, \Theta).$$

In the above definition Θ represents the assumed distribution of Q while Δ_Q is the distribution of Q observed from the data. The term $D(\Delta_Q, \Theta)$ measures the degree to which Δ_Q diverges from the prior distribution Θ, and $D(\Delta_{PQ}, \Delta_P \times \Theta)$ measures how far Δ_{PQ} diverges from the joint distribution of P and Q in case they were independent, and Q was distributed according to Θ.

The justification for the correcting term $D(\Delta_Q, \Theta)$ is given in the following theorem:

Theorem 2 *If P and Q are independent, and D is a Cziszar measure of divergence then $\Upsilon_{D,\Theta}(P \to Q) = 0$.*

Proof. In this case $\Delta_{PQ} = \Delta_P \times \Delta_Q$, and by Theorem 1 we have $\Upsilon_{D,\Theta}(P \to Q) = D(\Delta_P \times \Delta_Q, \Delta_P \times \Theta) - D(\Delta_Q, \Theta) = D(\Delta_Q, \Theta) - D(\Delta_Q, \Theta) = 0$. (Q.E.D.)

Note that if D is a Cziszar divergence $D = D_\phi$, then the invariance of these divergences implies:

$$\Upsilon_{D_\phi,\Theta}(P \to Q) = D_\phi(\Delta_{PQ}, \Delta_P \times \Theta) - D_\phi(\Delta_P \times \Delta_Q, \Delta_P \times \Theta).$$

3 Properties of the General Measure of Interestingness

Initially, we discuss several basic properties of the proposed measure.

Theorem 3 *If D is a Cziszar divergence, then*

$$\Upsilon_{D,\Delta_Q}(P \to Q) = \Upsilon_{D,\Delta_P}(Q \to P)$$

Proof. We have $\Upsilon_{D,\Delta_Q}(P \to Q) = D(\Delta_{PQ}, \Delta_P \times \Delta_Q)$, $\Upsilon_{D,\Delta_P}(Q \to P) = D(\Delta_{QP}, \Delta_Q \times \Delta_P)$, and the proof follows from the permutational symmetry of Cziszar's divergence [6]. (Q.E.D.)

The above property means that when the assumed distribution of the consequent is kept equal to the distribution observed from data, then the measure is symmetric with respect to the direction of the rule, i.e. exchanging the antecedent and consequent does not change the value of the interestingness.

Theorem 4 *Let D be a Cziszar divergence. If R is a set of attributes independent of P, and jointly of PQ, then, for any Θ*

$$\Upsilon_{D,\Theta}(RP \to Q) = \Upsilon_{D,\Theta}(P \to Q).$$

If R is a set of attributes independent of Q, and jointly of PQ, then

$$\Upsilon_{D,\Delta_{RQ}}(P \to RQ) = \Upsilon_{D,\Delta_Q}(P \to Q).$$

Proof.

$$\begin{aligned}
\Upsilon_{D,\Theta}(RP \to Q) &= D(\Delta_{RPQ}, \Delta_{RP} \times \Theta) - D(\Delta_Q, \Theta) \\
&= D(\Delta_R \times \Delta_{PQ}, \Delta_R \times \Delta_P \times \Theta) - D(\Delta_Q, \Theta) \\
&= D(\Delta_{PQ}, \Delta_P \times \Theta) - D(\Delta_Q, \Theta) = \Upsilon_{D,\Theta}(P \to Q) \\
&\quad \text{(by Theorem 1)}
\end{aligned}$$

For the second part, it follows from theorem 3 and from the first part of this theorem that

$$\begin{aligned}
\Upsilon_{D,\Delta_{RQ}}(P \to RQ) &= \Upsilon_{D,\Delta_P}(RQ \to P) = \Upsilon_{D,\Delta_P}(Q \to P) \\
&= \Upsilon_{D,\Delta_Q}(P \to Q).
\end{aligned}$$

(Q.E.D.)

The previous result gives a desirable property of $\Upsilon_{D,\Theta}$ since adding independent attributes should not affect rule's interestingness.

Note that if $\Theta = \Delta_Q$, that is, when Θ equals the observed distribution of the consequent, then Υ becomes symmetric and is not affected by adding independent attributes to either the antecedent or the consequent.

Next, we consider several important special cases of the interestingness measure.

If the divergence D and the assumed distribution used in the definition of the interestingness measure are chosen appropriately, then the interestingness $\Upsilon_{D,\Theta}(P \to Q)$ is proportional to a gain of the set of attributes of the consequent Q of the rule relative to the antecedent P. Both the Gini gain, $\mathsf{gain}_{\text{gini}}(Q, P)$, and the entropy gain, $\mathsf{gain}_{\text{shannon}}(Q, P)$, can be obtained by appropriate choice of D. Moreover a measure proportional to the χ^2 statistic can be obtained in that way.

Suppose that the attribute sets P, Q have the distributions

$$\Delta_P = \begin{pmatrix} x_1 & \cdots & x_m \\ p_1 & \cdots & p_m \end{pmatrix}, \text{ and } \Delta_Q = \begin{pmatrix} y_1 & \cdots & y_n \\ q_1 & \cdots & q_n \end{pmatrix}.$$

Let $\rho_{ij} = \{t \in \rho | t[P] = x_i \text{ and } t[Q] = y_j\}$ and let $p_{ij} = \frac{|\rho_{ij}|}{|\rho|}$ for $1 \leq i \leq m$ and $1 \leq j \leq n$.

Theorem 5 *Let $P \longrightarrow Q$ be a rule in the table $\tau = (T, H, \rho)$. If $D = D_{\mathrm{KL}}$ then*

$$\Upsilon_{D,\Theta}(P \to Q) = \mathsf{gain}_{\mathrm{shannon}}(Q, P),$$

regardless of the choice of Θ.

Proof. The definition of the Kullback-Leibler divergence allows us to write:

$$\begin{aligned}
\Upsilon_{D_{\mathrm{KL}},\Theta}(P \to Q) &= D_{\mathrm{KL}}\left(\Delta_{PQ}, \Delta_P \times \Theta\right) - D_{\mathrm{KL}}\left(\Delta_Q, \Theta\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{p_i \theta_j} - \sum_{j=1}^{n} q_j \log \frac{q_j}{\theta_j} \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} \log p_{ij} - \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} \log p_i \\
&\quad - \sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} \log \theta_j - \sum_{j=1}^{n} q_j \log q_j + \sum_{j=1}^{n} q_j \log \theta_j \\
&= \mathcal{H}(P) + \mathcal{H}(Q) - \mathcal{H}(PQ) = \mathsf{gain}_{\mathrm{shannon}}(Q, P), \\
&\quad \text{(by Equality (2))}
\end{aligned}$$

which completes the proof. (Q.E.D.)

The above theorem means that for the case D_{KL} the family of measures generated by Θ reduces to a single measure: the Shannon gain (mutual information). This is not the case for other divergences.

Theorem 6 *Let $P \longrightarrow Q$ be a rule in the table $\tau = (T, H, \rho)$. If $D = D_{\chi^2}$ and $\Theta = \mathcal{U}_n$, where $n = |\mathrm{dom}(Q)|$, then*

$$\Upsilon_{D,\Theta}(P \to Q) = n \cdot \mathsf{gain}_{\mathrm{gini}}(Q, P).$$

Proof. We have

$$\begin{aligned}
\Upsilon_{\chi^2,\mathcal{U}_n}(P \to Q) &= D_{\chi^2}\left(\Delta_{PQ}, \Delta_P \times \mathcal{U}_n\right) - D_{\chi^2}\left(\Delta_Q, \mathcal{U}_n\right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_{ij}^2}{\frac{p_i}{n}} - \sum_{j=1}^{n} \frac{q_j^2}{\frac{1}{n}} \\
&= n\left(\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{p_{ij}^2}{p_i} - \sum_{j=1}^{n} q_j^2\right) \\
&= n \cdot \mathsf{gain}_{\mathrm{gini}}(Q, P), \\
&\quad \text{(by Equality (3))}
\end{aligned}$$

which is the desired equality. (Q.E.D.)

Theorem 7 *We have $\Upsilon_{D_{\chi^2},\Delta_Q}(P \longrightarrow Q)$ is proportional to $\chi^2(P, Q)$, the chi-squared statistics [1] for attribute sets P, Q.*

Proof.

$$\begin{aligned}
\Upsilon_{D_{\chi^2},\Delta_Q}(P \longrightarrow Q) &= D_{\chi^2}\left(\Delta_{PQ}, \Delta_P \times \Delta_Q\right) - D_{\chi^2}\left(\Delta_Q, \Delta_Q\right) \\
&= D_{\chi^2}\left(\Delta_{PQ}, \Delta_P \times \Delta_Q\right) \\
&= \frac{\chi^2(P, Q)}{|\rho|}
\end{aligned}$$

(Q.E.D.)

Note that above we treat attribute sets $P = \{A_1, \ldots, A_r\}$ and $Q = \{B_1, \ldots, B_s\}$ as single attributes with the domains given by (1). This is appropriate, since we are interested in how one *set* of attributes P influences another *set* of attributes Q. Another way, used in [3], is to compute $\chi^2(A_1, \ldots, A_r, B_1, \ldots, B_s)$, however this is not what we want.

The case when $D = D_{\chi^2}$ is of practical interest since it includes two widely used measures (χ^2, and $\mathsf{gain}_{\mathrm{gini}}$) as special cases, and allows for obtaining a continuum of measures "in between" the two.

Theorem 8 proven below shows that the generalized measure interestingness $\Upsilon_{D,\Theta}(P \to Q)$ is minimal when P and Q are independent and thus, it justifies our definition of this measure through variational considerations. We begin with a technical result.

Lemma 1 *Let $P = (p_{ij}) \in \mathbb{R}^{n \times n}$ be an $n \times n$-matrix with non-negative entries such that $\sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} = 1$. For $\eta = (1, \ldots, 1) \in \mathbb{R}^n$ let $p^T = P\eta^T \in \mathbb{R}^n$ and $q = \eta P \in \mathbb{R}^n$.*

If P can be written as $P = u^T v$, for some $u, v \in \mathbb{R}^n$, then $P = p^T q$.

Proof. Since $P = u^T v$, we have $p = \eta P^T = \eta v^T u = (v\eta^T)u$, and $q = \eta u^T v = (\eta u^T)v$. If a, b are the non-negative numbers $a = v\eta^T$ and $b = \eta u^T$, then we have $p = au$ and $q = bv$. Note that the condition $\sum_{i=1}^{n} \sum_{j=1}^{n} p_{ij} = 1$ can be written as $\eta P \eta^T = 1$, or as $\eta u^T v \eta^T = 1$. Using the associativity of the matrix product we have $ab = \eta u^T v \eta^T = 1$, so we can write $P = u^T v = abu^T v = (au)^T (bv) = p^T q$, which is the desired equality. (Q.E.D.)

Theorem 8 *Let $\Upsilon_{D,\Theta}$ be the measure of interestingness generated by the assumed distribution Θ and the Kullback-Leibler divergence, or the χ^2-divergence and let $P \to Q$ be a rule. For any fixed attribute distribution Δ_P, Δ_Q and a fixed distribution Θ, the value of $\Upsilon_{D,\Theta}(P \to Q)$ is minimal (and equal to 0) if only if $\Delta_{PQ} = \Delta_P \times \Delta_Q$, i.e., when P and Q are independent.*

Proof. It is clear that if P and Q are independent, then we have in both cases $\Upsilon_{D,\Theta}(P \to Q) = 0$.

When $D = D_{\mathrm{KL}}$ the result follows from the properties of Shannon gain/mutual information [7].

We need to prove the result for $D = D_{\chi^2}$. It has been noted in Chapter 1 that D_{χ^2} is a special case of Cziszar divergence for $\phi(x) = x^2 - x$. From the conditions on ϕ it follows that for all Cziszar's divergences the respective functions ϕ have the property that the inverses of their first derivatives are monotonic functions and therefore can be inverted. Indeed, in the case of D_{χ^2} we have $\phi(x) = x^2 - x$, and $(\phi')^{-1}(x) = x/2 + 1/2$.

We will use *Lagrange multipliers* method to find the minimum of $D_{\chi^2}(\Delta_{PQ}, \Delta_P \times \Theta)$ subject to the following set of constraints:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} = 1 \tag{4}$$

$$\sum_{j=1}^{n} p_{ij} = p_i \tag{5}$$

$$\sum_{i=1}^{m} p_{ij} = q_j \tag{6}$$

The Lagrangian is

$$\begin{aligned} L &= \sum_{i=1}^{m}\sum_{j=1}^{n} p_i\theta_j\phi\left(\frac{p_{ij}}{p_i\theta_j}\right) + \lambda\left(\sum_{i=1}^{m}\sum_{j=1}^{n} p_{ij} - 1\right) \\ &+ \sum_{i=1}^{m}\lambda_i\left(\sum_{j=1}^{n} p_{ij} - p_i\right) + \sum_{j=1}^{n}\mu_j\left(\sum_{i=1}^{m} p_{ij} - q_j\right), \end{aligned}$$

and

$$\frac{\partial L}{\partial p_{ij}} = \phi'\left(\frac{p_{ij}}{p_i\theta_j}\right) + \lambda + \lambda_i + \mu_j. \tag{7}$$

By equating (7) to zero we get:

$$\phi'\left(\frac{p_{ij}}{p_i\theta_j}\right) = -(\lambda + \lambda_i + \mu_j).$$

In the case of the D_{χ^2} measure of divergence we have $(\phi')^{-1}(x) = \frac{x}{2} + \frac{1}{2}$. Therefore, p_{ij} can be written as

$$p_{ij} = p_i\theta_j\left[\frac{-\lambda - \lambda_i - \mu_j}{2} + \frac{1}{2}\right] = \frac{1}{2}p_i\theta_j(1 - \lambda - \lambda_i - \mu_j).$$

Substituting into (5) we get $\sum_{j=1}^{n} p_{ij} = \frac{1}{2}\sum_{j=1}^{n} p_i\theta_j(1-\lambda-\lambda_i-\mu_j) = p_i$, and

$$\sum_{j=1}^{n}\theta_j(1 - \lambda - \lambda_i - \mu_j) = 2.$$

After splitting the sum we get $1 - \lambda - \lambda_i - \sum_{j=1}^{n}\theta_j\mu_j = 2$, and

$$\lambda_i = -\lambda - \sum_{j=1}^{n}\theta_j\mu_j - 1 = c_\alpha.$$

Similarly, substituting into (6) we get $\sum_{i=1}^{m} p_{ij} = \frac{1}{2}\sum_{i=1}^{m} p_i\theta_j(1-\lambda-\lambda_i-\mu_j) = q_j$, and

$$\sum_{i=1}^{m} p_i(1 - \lambda - \lambda_i - \mu_j) = 2\frac{q_j}{\theta_j}.$$

After splitting the sum we get $1 - \lambda - \mu_j - \sum_{i=1}^{m} p_i \lambda_i = 2\frac{q_j}{\theta_j}$, and

$$\mu_j = 1 - \lambda - \sum_{i=1}^{m} p_i \lambda_i - 2\frac{q_j}{\theta_j} = c_\beta - 2\frac{q_j}{\theta_j}.$$

Thus,

$$p_{ij} = \frac{1}{2} p_i \theta_j (1 - \lambda - c_\alpha - c_\beta + 2\frac{q_j}{\theta_j}) = \frac{1}{2} p_i \theta_j (c_\gamma + 2\frac{q_j}{\theta_j}),$$

for some constant c_γ, which means that the matrix P has the form $P = u^T v$ for some $u, v \in \mathbb{R}^n$. By Lemma 1, we have $P = p^T q$. (Q.E.D.)

We proved that $\mathsf{gain}_{\text{shannon}}$ and $\mathsf{gain}_{\text{gini}}$ are equivalent to $\Upsilon_{D_{\text{KL}},\mathcal{U}_n}$ and $\Upsilon_{D_{\chi^2},\mathcal{U}_n}$, respectively. It is thus natural to define a notion of *gain* for any divergence D as

$$\mathsf{gain}_D(P \to Q) = \Upsilon_{D,\mathcal{U}_n}(P \to Q).$$

Let $\Delta_Q|p_i$ denote the probability distribution of Q conditioned on $P = p_i$. For any Cziszar measure D_ϕ we have:

$$\begin{aligned}
\mathsf{gain}_{D_\phi}(P \to Q) &= D_\phi(\Delta_{PQ}, \Delta_P \times \mathcal{U}_n) - D_\phi(\Delta_Q, \mathcal{U}_n) \\
&= \sum_{i=1}^{m} p_i \sum_{j=1}^{n} \frac{1}{n} \phi\left(\frac{p_{ij}}{p_i \cdot \frac{1}{n}}\right) - D_\phi(\Delta_Q, \mathcal{U}_n) \\
&= -\left[D_\phi(\Delta_Q, \mathcal{U}_n) - \sum_{i=1}^{m} p_i D_\phi(\Delta_Q|p_i, \mathcal{U}_n)\right].
\end{aligned}$$

As special cases $\mathsf{gain}_{\text{gini}} \equiv \mathsf{gain}_{\chi^2}$, and $\mathsf{gain}_{\text{shannon}} \equiv \mathsf{gain}_{\text{KL}}$.

A parameterized version of Υ that takes into account the degree of confidence in the distribution of the consequent as it results from the data is introduced next.

Let us define the probability distribution $\Theta_a, a \in [0, 1]$ by

$$\Theta_a = a\Delta_Q + (1 - a)\mathcal{U}_n.$$

The value of a expresses the amount of confidence we have in Δ_Q estimated from the data. The value $a = 1$ means total confidence, we assume the probability estimated from data as the true probability distribution of Q. On the other hand, $a = 0$ means that we have no confidence in the estimate and use some prior distribution of Q instead. In our case, the prior is the uniform distribution $\mathcal{U}_n$. Note that $\Theta_1 = \Delta_Q$, and $\Theta_0 = \mathcal{U}_n$.

So, Θ_a is the *a posteriori* distribution for Q.

We can now define

$$\Upsilon_{D,a} = \Upsilon_{D,\Theta_a}.$$

Note that when $D = D_{\chi^2}$, we have (up to a constant factor) both $\chi^2(P \to Q)$ and $\text{gini}_{\mathsf{gain}}(P \to Q)$ as special cases of $\Upsilon_{D_{\chi^2},a}$. Moreover by taking different values of parameter a we can obtain a continuum of measures in between the two.

As noted before, both D_{χ^2} and D_{KL} divergence measures are special cases of Havrda-Charvát divergence $D_{\mathcal{H}_\alpha}$ for $\alpha \to 1$, and $\alpha = 2$ respectively. We can thus introduce $\Upsilon_{\alpha,a} = \Upsilon_{D_{\mathcal{H}_\alpha},\Theta_a}$, which allows us to obtain a family of interestingness measures, including (up to a constant factor) all three measures given in Section 3 as special cases, by simply changing two real valued parameters α and a.

Also note that for $a = 0$, we obtain a family of gains (as defined in chapter 3) for all the Havrda-Charvát divergences.

4 Experimental results

We evaluated the new measure on a simple synthetic dataset and on data from the UCI machine learning repository [2]. We concentrated on the case $D = D_{\chi^2}$, as potentially most useful in practice, and found interestingness of rules for different values of parameter a (see chapter 3)

4.1 Synthetic data

To ensure measures throughout the family handle obvious cases correctly, and to make it easy to observe properties of the measure for different values of parameter a we first evaluated the rules on a synthetic dataset with 3 attributes A, B, C and with known probabilistic dependencies between them.

Values of attributes A and B have been generated from known probability distributions:

$$\Delta_A = \begin{pmatrix} 0 & 1 & 2 \\ 0.1 & 0.5 & 0.4 \end{pmatrix}, \Delta_B = \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \end{pmatrix}.$$

Attribute C depends on attribute A. Denote $\Delta_C|i$ the distribution of C conditioned upon $A = i$. We used

$$\Delta_C|0 = \begin{pmatrix} 0 & 1 \\ 0.2 & 0.8 \end{pmatrix}, \Delta_C|1 = \begin{pmatrix} 0 & 1 \\ 0.5 & 0.5 \end{pmatrix}, \Delta_C|2 = \begin{pmatrix} 0 & 1 \\ 0.7 & 0.3 \end{pmatrix},$$

One million data points have been generated according to this distribution, for a few values of a we sorted all possible rules based on their $\Upsilon_{D_{\chi^2},a}$ interestingness values. Results are given in Table 4.1.

Discussion

1. Attribute B is totally independent of both A and C, so any rule containing only B as the antecedent or consequent should have interestingness 0. The experiments confirm this, for all values of parameter a such rules have interestingness close to zero, significantly lower than the interestingness of any other rules.
2. For $a = 0$ (the first quarter of the table) Υ becomes the Gini gain, a measure that is strongly asymmetric (and could thus suggest the direction of the dependence) and strongly affected by adding extra independent attributes to the consequent (which is undesirable).

3. For $a = 1$ (the last quarter of the table) Υ becomes (up to a constant factor) the χ^2 measure of dependence. This measure is totally symmetric and not affected by presence of independent attributes in either antecedent or consequent. Indeed, it can be seen that all rules involving A and C have the same interestingness regardless of the presence of B in the antecedent or consequent.

4. As a varies from 0 to 1 the intermediate measures can be seen to become more and more symmetric. Measures for a being close to but less than 1 could be of practical interest since they seem to 'combine the best of the two worlds', that is, are still asymmetric and pretty insensitive to presence of independent attributes in the consequent. E.g. for $a = 0.9$ all rules having A in the antecedent and C in the consequent have interestingness close to 0.09, while rules having C in the antecedent and A in the consequent have all interestingness close to 0.082 regardless of the presence or absence of B in the consequents. So for $a = 0.9$ the intermediate measure correctly ranked the rules indicating the true direction of the relationship.

4.2 The mushroom database

We then repeated the above experiment on data from the UCI machine learning repository [2]. Here we present results for the *agaricus-lepiota* database containing data on North American Mushrooms. To make the ruleset size manageable we restrict ourselves to rules involving the *class* attribute indicating whether the mushroom is edible or poisonous.

In the experiment we enumerated all rules involving up to 3 attributes and ranked them by interestingness for different values of parameter a. Top ten rules for each value of a are shown in Table 4.1. For $a = 1$ the symmetric rules were removed.

We noticed that for any value of a most of the rules involve the *odor* attribute. Indeed the inspection of data revealed that knowing the mushroom's odor allows for identifying its class with 98.5% accuracy, far better than for any other attribute.

We note also that similar rules are ranked close to the top for all values of a, which proves that measures thoughout the family identify dependencies correctly. From data omitted in the tables it can be observed that, as in the case of synthetic data, when a approaches 1 the measures become more and more symmetric and unaffected by independent attributes in the consequent.

4.3 Conclusions

It has been shown experimentally that measures throughout the Υ family are useful for discovering interesting dependencies among data attributes. By modifying a numerical attribute we can obtain a whole spectrum of measure of varying degree of symmetry and dependence on the presence of extra attributes in the rule consequent. Especially interesting seem to be measures with a parameter close to, but less than 1, which combine the

rule	$\Upsilon_{D_{\chi^2},0}$	rule	$\Upsilon_{D_{\chi^2},0.5}$
$A \to BC$	0.122061	$A \to BC$	0.0989161
$C \to AB$	0.0896776	$AB \to C$	0.0898611
$AB \to C$	0.0896287	$A \to C$	0.089861
$A \to C$	0.0896287	$C \to AB$	0.0769886
$BC \to A$	0.065851	$BC \to A$	0.0683164
$C \to A$	0.0658484	$C \to A$	0.0683142
$B \to AC$	3.16585e-06	$B \to AC$	2.50502e-06
$B \to A$	2.7369e-06	$B \to A$	2.35091e-06
$AC \to B$	1.37659e-06	$AC \to B$	1.51849e-06
$A \to B$	1.32828e-06	$A \to B$	1.46355e-06
$B \to C$	1.70346e-07	$B \to C$	1.72781e-07
$C \to B$	1.10069e-07	$C \to B$	1.22814e-07

rule	$\Upsilon_{D_{\chi^2},0.9}$	rule	$\Upsilon_{D_{\chi^2},1}$
$A \to BC$	0.0908769	$BC \to A$	0.0905673
$AB \to C$	0.0903859	$A \to BC$	0.0905673
$A \to C$	0.0903859	$C \to AB$	0.0905654
$C \to AB$	0.0834734	$AB \to C$	0.0905654
$BC \to A$	0.082009	$A \to C$	0.0905653
$C \to A$	0.082007	$C \to A$	0.0905653
$B \to AC$	2.19739e-06	$AC \to B$	2.15872e-06
$B \to A$	2.12646e-06	$B \to AC$	2.15872e-06
$AC \to B$	1.95101e-06	$A \to B$	2.08117e-06
$A \to B$	1.87986e-06	$B \to A$	2.08017e-06
$B \to C$	1.73782e-07	$C \to B$	1.74126e-07
$C \to B$	1.57306e-07	$B \to C$	1.74126e-07

Table 1: Rules on synthetic data ordered by $\Upsilon_{D_{\chi^2},a}$ for different values of a.

rule	$\Upsilon_{D_{\chi^2},0}$
class→odor ring-type	9.84024
class→odor spore-print-color	9.16709
class→odor veil-color	8.22064
class→odor gill-attachment	8.2026
class→gill-color spore-print-color	7.82161
class→ring-type spore-print-color	7.62564
class→odor stalk-root	7.60198
class→gill-color ring-type	7.28972
class→odor stalk-color-above-ring	7.19584
class→odor stalk-color-below-ring	7.14197

rule	$\Upsilon_{D_{\chi^2},0.9}$
odor→class stalk-root	3.61877
class stalk-root→odor	3.2782
odor→class cap-color	2.59777
odor→class ring-type	2.54896
odor→class spore-print-color	2.54864
stalk-color-above-ring→class stalk-color-below-ring	2.47669
class cap-color→odor	2.46105
odor→class gill-color	2.45027
stalk-color-below-ring→class stalk-color-above-ring	2.38593
class spore-print-color→odor	2.35384

rule	$\Upsilon_{D_{\chi^2},1}$
class stalk-root→odor	4.11701
class stalk-color-below-ring→stalk-color-above-ring	3.38287
stalk-color-below-ring→class stalk-color-above-ring	3.37968
class ring-type→odor	2.98764
class cap-color→odor	2.85308
odor→class gill-color	2.82423
odor→class spore-print-color	2.56331
odor→class stalk-color-below-ring	2.44004
class stalk-color-above-ring→odor	2.42725
class gill-color→spore-print-color	2.42224

Table 2: Rules on mushroom dataset ordered by $\Upsilon_{D_{\chi^2},a}$ for different values of a.

relative robustness against extra independent attributes, while retaining the asymmetry suggesting the direction of the dependence.

5 Open Problems and Future Directions

Above we assumed complete confidence in the estimate of the distribution of P from the data. We may want to relax this restriction and assume that P has some assumed distribution Ψ (not necessarily equal to Δ_P), and Q the prior distribution Θ. We can then generalize Υ as follows

$$\Upsilon'_{D,\Theta,\Psi}(P \to Q) = D(\Delta_{PQ}, \Psi \times \Theta) - D(\Delta_Q, \Theta) - D(\Delta_P, \Psi).$$

When $\Psi = \Delta_P$, Υ' reduces to Υ defined above. Some of the properties of Υ are preserved by this new definition. For example, if $D = D_{\mathrm{KL}}$, and P, Q be independent, then $\Upsilon'_{\Theta,\Psi,D}(P \to Q) = 0$. Also, if $P \longrightarrow Q$ is a rule in the table $\tau = (T, H, \rho)$ and $D = D_{\mathrm{KL}}$ then $\Upsilon'_{D,\Theta,\Psi}(P \to Q) = \mathsf{gain}_{\mathrm{shannon}}(Q, P)$ regardless of the choice of Θ and Ψ.

Further theoretical and experimental evaluation of the new measure is necessary. It might be of practical interest to modify the generalized definition of gain so that, being asymmetric, it is not affected by adding independent attributes to the consequent.

It would also be of practical significance to generalize the measure to express the interestingness of a rule with respect to a system of beliefs (that could be represented for example by a set of rules). Then, the rule would be considered interesting if its probability distribution would be significantly different from the one expected based on the set of beliefs. See [12] for a discussion of a similar problem.

References

[1] Bayardo R.J. and R. Agrawal, *Mining the Most Interesting Rules*, Proc. of the 5th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, pp. 145-154, August 1999.

[2] Blake C.L., and Merz C.J. *UCI Repository of machine learning databases* [`http://www.ics.uci.edu/~mlearn/MLRepository.html`] Irvine, CA: University of California, Dept. of Information and Computer Science.

[3] Silverstein C., Brin S. and Motwani R., *Beyond Market Baskets: Generalizing Association Rules to Dependence Rules* Data Mining and Knowledge Discovery, 2(1998), pp. 39-68

[4] Cziszar I., *A Class of Measures of Informativity of Observation Channels*, Periodic Math. Hungarica, 2:191-213, 1972.

[5] Havrda J.H., Charvát F., *Quantification Methods of Classification Processes: Concepts of Structural α Entropy*, Kybernetica, 3:30-35, 1967.

[6] Kapur J.N. and Kesavan H.K., *Entropy Optimization Principles with Applications*, Academic Press, San Diego, 1992.

[7] McEliece R.J., *The Theory of Information and Coding. A mathematical Framework for Communication*, Encyclopedia of Mathematics and its Applications, Addisson-Wesley, Reading Massachusetts, 1977.

[8] Kvalseth T.O., *Entropy and Correlation: Some comments*, IEEE Trans. on Systems, Man and Cybernetics, SMC-17(3):517–519.

[9] Mitchell T.M.. *Machine Learning*, McGraw-Hill, ISBN: 0070428077.

[10] Morimoto Y., Fukuda T., Matsuzawa H., Tokuyama T. and Yoda K. *Algorithms for Mining Association Rules for Binary Segmentations of Huge Categorical Databases*, Proc. of the 24th Conf. on Very Large Databases, pp. 380-391, 1998

[11] Morishita S., *On Classification and Regression* Proc. of the First Int'l Conf. on Discovery Science — Lecture Notes in Artificial Intelligence 1532:40–57, 1998

[12] Padmanabhan B. and Tuzhilin A. *Unexpectedness as a measure of interestingness in knowledge discovery* Decision and Support Systems 27(1999), pp. 303-318

[13] Simovici, D. A. and Tenney R. L. *Relational Database Systems*, Academic Press, 1995, San Diego.

[14] Wehenkel L., *On uncertainty Measures Used for Decision Tree Induction*, Info. Proc. and Manag. of Uncertainty in Knowledge-Based Systems (IPMU'96), July 1-5, 1996, Granada Spain, pp. 413–418.