A General Measure of Rule Interestingness

Szymon Jaroszewicz and Dan A. Simovici

Department of Computer Science
University of Massachusetts at Boston
Motivation:

- Data Mining algorithms produce 10s of thousands of rules
- Need to assess rules quality, need for measures of interestingness
- Several measures are used: entropy gain, gini gain, chi squared

**Our work:** A new measure or rule interestingness $\Upsilon$ generalizing the 3 above.
Rule:

\[ P \rightarrow Q \]

where \( P \) and \( Q \) are sets of attributes.

What we know about the rule:

- Estimate of joint distribution \( \Delta_P = (p_i) \) of \( P \)
- Estimate of joint distribution \( \Delta_Q = (q_j) \) of \( Q \)
- Estimate of joint distribution \( \Delta_{PQ} = (p_{ij}) \) of \( PQ \)

Different from association rules where only some of the probabilities are known.
Examples of measures of interestingness using full probability distributions:

1. entropy gain

\[
\text{gain}_{\text{shannon}}(P \rightarrow Q) = - \sum_{j=1}^{n} q_j \log q_j + \sum_{i=1}^{m} \sum_{j=1}^{n} p_{ij} \log \frac{p_{ij}}{q_j}
\]

2. gini gain

\[
\text{gain}_{\text{gini}}(P \rightarrow Q) = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{p_{ij}^2}{q_j} - \sum_{i=1}^{m} p_i^2
\]

3. Chi squared

\[
\chi^2(P \rightarrow Q) = |\rho| \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{(p_{ij} - p_i q_j)^2}{p_i q_j}
\]
Notion of divergence (distance) between two probability distributions $\Delta = (p_1, p_2, \ldots, p_n)$, and $\Delta' = (q_1, q_2, \ldots, q_n)$

- **Kullback-Leibler divergence (cross-entropy)**

\[
D_{KL}(\Delta, \Delta') = \sum_{i=1}^{n} p_i \log \frac{p_i}{q_i}
\]

- **$\chi^2$ divergence**

\[
D_{\chi^2}(\Delta, \Delta') = \sum_{i=1}^{n} \frac{(p_i - q_i)^2}{q_i}
\]
Rule: \( P \rightarrow Q \)

- Assume \( \Delta_P \) estimated from data is the true distribution of \( P \)
- Uniform prior distribution \( \mathcal{U} \) of \( Q \)
- Laplace estimate for \textit{a posteriori} distribution \( \Theta \) of \( Q \), where

\[
\Theta = \frac{|\rho| \Delta_Q + M \mathcal{U}}{|\rho| + M}
\]

\( M = 0 \) total confidence in the estimate
\( M \rightarrow \infty \) no confidence, use the apriori distribution

- To avoid limits, denote \( a = \frac{|\rho|}{|\rho| + M} \), and

\[
\Theta_a = a \Delta_Q + (1 - a) \mathcal{U}
\]

\( a = 1 \) total confidence in the estimate
\( a = 0 \) no confidence, use the apriori distribution
Rule: $P \rightarrow Q$

- $\Delta_P$ the distribution of $P$
- $\Theta_a$ a posteriori distribution of $Q$ depending on the degree $a$ of confidence in the data

**Assumptions:**

1. The more $P$ and $Q$ depend on each other the more interesting the rule. Use distribution divergence $D$ to measure dependence:

   $$D(\Delta_{PQ}, \Delta_P \times \Theta_a)$$

2. When $P$ and $Q$ are independent, interestingness should be 0

**Our measure of interestingness:**

$$\Upsilon_{D,a}(P \rightarrow Q) = D(\Delta_{PQ}, \Delta_P \times \Theta_a) - D(\Delta_Q, \Theta_a)$$
Special cases

**Entropy gain** $D = D_{KL}$, any value of $a$

\[
\Upsilon_{D_{KL},a}(P \rightarrow Q) = \text{gain}_{\text{shannon}}(P \rightarrow Q) = D_{KL}(\Delta_{PQ}, \Delta_P \times \Delta_Q) = \text{mutual information}(P, Q)
\]

**Gini gain** $D = D_{\chi^2}$, $a = 0$ (no confidence in estimate of $\Delta_Q$)

\[
\Upsilon_{D_{\chi^2},0}(P \rightarrow Q) \propto \text{gain}_{\text{gini}}(P \rightarrow Q)
\]

**Chi squared** $D = D_{\chi^2}$, $a = 1$ (total confidence in estimate of $\Delta_Q$)

\[
\Upsilon_{D_{\chi^2},1}(P \rightarrow Q) \propto \chi^2(P \rightarrow Q)
\]

For $a \in [0, 1]$ we obtain a continuum of measures between gain_{gini} and $\chi^2$
Properties of intermediate measures

- For any value of \( a \in [0, 1] \),

\[
\Upsilon_{D_{\chi^2}, a}(P \rightarrow Q) \geq 0
\]

with equality iff \( P \) and \( Q \) are independent.

- \( R \) is a set of attributes independent of \( P \) and \( Q \)
  
  For any value of \( a \in [0, 1] \)

\[
\Upsilon_{D_{\chi^2}, a}(PR \rightarrow Q) = \Upsilon_{D_{\chi^2}, a}(P \rightarrow Q)
\]

For \( a = 1 \)

\[
\Upsilon_{D_{\chi^2}, a}(P \rightarrow QR) = \Upsilon_{D_{\chi^2}, a}(P \rightarrow Q)
\]

- Independent attributes in \( P \) do not affect interestingness.
  Generally not true about \( Q \).
<table>
<thead>
<tr>
<th>$a = 1$</th>
<th>$a = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>symmetric (unconditional)</td>
<td>asymmetric (conditional)</td>
</tr>
<tr>
<td>not affected by</td>
<td>affected by independent</td>
</tr>
<tr>
<td>independent attributes</td>
<td>attributes in consequent</td>
</tr>
</tbody>
</table>

**Why use intermediate measures?**

Choosing value of $a$ close to (but less than) 1

- Asymmetric, may suggest the direction of dependence
- Affected in a very small degree by independent attributes
Synthetic dataset 3 attributes: $A \rightarrow C, B$

Probability distributions:

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

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

$B$ independent of $A, C$ and jointly of $AC$
Rules from the synthetic dataset sorted by $\gamma_{\chi^2}$ for various $a$

<table>
<thead>
<tr>
<th>rule</th>
<th>$\gamma_{D_{\chi^2},0}$</th>
<th>rule</th>
<th>$\gamma_{D_{\chi^2},0.9}$</th>
<th>rule</th>
<th>$\gamma_{D_{\chi^2},1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A \rightarrow BC$</td>
<td>0.122</td>
<td>$A \rightarrow BC$</td>
<td>0.090</td>
<td>$BC \rightarrow A$</td>
<td>0.090</td>
</tr>
<tr>
<td>$C \rightarrow AB$</td>
<td>0.090</td>
<td>$AB \rightarrow C$</td>
<td>0.090</td>
<td>$A \rightarrow BC$</td>
<td>0.090</td>
</tr>
<tr>
<td>$AB \rightarrow C$</td>
<td>0.090</td>
<td>$A \rightarrow C$</td>
<td>0.090</td>
<td>$C \rightarrow AB$</td>
<td>0.090</td>
</tr>
<tr>
<td>$A \rightarrow C$</td>
<td>0.090</td>
<td>$C \rightarrow AB$</td>
<td>0.083</td>
<td>$AB \rightarrow C$</td>
<td>0.090</td>
</tr>
<tr>
<td>$BC \rightarrow A$</td>
<td>0.065</td>
<td>$BC \rightarrow A$</td>
<td>0.082</td>
<td>$A \rightarrow C$</td>
<td>0.090</td>
</tr>
<tr>
<td>$C \rightarrow A$</td>
<td>0.065</td>
<td>$C \rightarrow A$</td>
<td>0.082</td>
<td>$C \rightarrow A$</td>
<td>0.090</td>
</tr>
<tr>
<td>$B \rightarrow AC$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow AC$</td>
<td>$\approx 0$</td>
<td>$AC \rightarrow B$</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>$B \rightarrow A$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow A$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow AC$</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>$AC \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$AC \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$A \rightarrow B$</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>$A \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$A \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow A$</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>$B \rightarrow C$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow C$</td>
<td>$\approx 0$</td>
<td>$C \rightarrow B$</td>
<td>$\approx 0$</td>
</tr>
<tr>
<td>$C \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$C \rightarrow B$</td>
<td>$\approx 0$</td>
<td>$B \rightarrow C$</td>
<td>$\approx 0$</td>
</tr>
</tbody>
</table>
The mushroom dataset (3 attribute rules)

| $\mathcal{Y}_{D_{\chi^2},0}$ | class $\rightarrow$ odor ring-type | 9.84 |
|                            | class $\rightarrow$ odor spore-print-color | 9.17 |
|                            | class $\rightarrow$ odor veil-color | 8.22 |
|                            | class $\rightarrow$ odor gill-attachment | 8.20 |
|                            | class $\rightarrow$ gill-color spore-print-color | 7.82 |

| $\mathcal{Y}_{D_{\chi^2},0.9}$ | odor $\rightarrow$ class stalk-root | 3.62 |
|                                | class stalk-root $\rightarrow$ odor | 3.28 |
|                                | odor $\rightarrow$ class cap-color | 2.60 |
|                                | odor $\rightarrow$ class ring-type | 2.55 |
|                                | odor $\rightarrow$ class spore-print-color | 2.55 |

| $\mathcal{Y}_{D_{\chi^2},1}$ | class stalk-root $\rightarrow$ odor | 4.12 |
|                              | class stalk-color-below-ring $\rightarrow$ stalk-color-above-ring | 3.38 |
|                              | stalk-color-below-ring $\rightarrow$ class stalk-color-above-ring | 3.38 |
|                              | class ring-type $\rightarrow$ odor | 2.99 |
|                              | class cap-color $\rightarrow$ odor | 2.85 |

* symmetric rules removed
Further generalizations

Using the Havrda-Charvát divergence

\[ D_{\mathcal{H}_\alpha} = \frac{1}{\alpha - 1} \left( \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} - 1 \right) \]

Special cases:
\( D_{\chi^2} \) is obtained when \( \alpha = 2 \)
\( D_{KL} \) is obtained when \( \alpha \to 1 \)

Define:

\[ \Upsilon_{\alpha,a}(P \to Q) = \Upsilon_{D_{\mathcal{H}_\alpha},a}(P \to Q) \]

This way by changing 2 parameters we can obtain entropy gain, gini gain, chi squared as special cases of a single general measure.
Intermediate measures

\[
\alpha \quad \text{gini gain} \quad a \quad \text{chi sq.}
\]

\[
\alpha \quad \text{ent. gain}
\]

intermediate measures
Prior distributions
Assume arbitrary prior distribution of $Q$ e.g. reflecting background knowledge.

Let $\Theta$ be a posteriori distribution of $Q$

The following hold

- For every distribution $\Theta$
  \[ \Upsilon_{D_{\chi^2},\Theta}(P \rightarrow Q) \geq 0 \]
  with equality iff $P$ and $Q$ are independent.

- For every distribution $\Theta$
  \[ \Upsilon_{D_{KL},\Theta}(P \rightarrow Q) = \text{gain}_{\text{shannon}}(P \rightarrow Q) \]
Prior distributions

Assume a prior/posterior distribution also on $P$:

$$\Upsilon_{D,\Theta,\Psi}(P \rightarrow Q) = D(\Delta_{PQ}, \Psi \times \Theta) - D(\Delta_Q, \Theta) - D(\Delta_P, \Psi).$$

Properties:

- For all distributions $\Theta, \Psi$

  $$\Upsilon_{D_{KL},\Theta,\Psi}(P \rightarrow Q) = \text{gain}_{\text{shannon}}(Q \rightarrow P)$$

- We cannot guarantee that if $P$ and $Q$ are independent, then $\Upsilon_{D_{KL},\Theta,\Psi}(P \rightarrow Q) = 0$. 
Further research

1. More experimental work is necessary
2. Apply the measure to decision tree induction
3. Investigate how the measure behaves if background knowledge is used as a prior for $Q$
4. More work on rule interestingness with respect to background knowledge