The Normalized χ^2 Measure for Association Rule Evaluation

Let C and A be two attributes with domains $\text{dom}(A) = \{a_1, \ldots a_{n_A}\}$ and $\text{dom}(C) = \{c_1, \ldots c_{n_C}\}$, respectively, and let $\mathcal{X}$ be a dataset over C and A. Let N_{ij}, $1 \le i \le n_C$, $1 \le j \le n_A$, be the number of sample cases in $\mathcal{X}$ that contain both the attribute values c_i and a_j. Furthermore, let

$$N_{i.} = \sum_{j=1}^{n_A} N_{ij}, \qquad N_{.j} = \sum_{i=1}^{n_C} N_{ij}, \qquad \text{and} \qquad N_{..} = \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} N_{ij} = |\mathcal{X}|.$$

Finally, let

$$p_{i.} = \frac{N_{i.}}{N_{..}}, \qquad p_{.j} = \frac{N_{.j}}{N_{..}}, \qquad \text{and} \qquad p_{ij} = \frac{N_{ij}}{N_{..}}$$

be the probabilities of the attribute values and their combinations, as they can be estimated from these numbers. Then the well-known χ^2 measure is usually defined as

$$\begin{aligned}
\chi^2(C, A) &= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{(E_{ij} - N_{ij})^2}{E_{ij}} \qquad \text{where} \quad E_{ij} = \frac{N_{i.} N_{.j}}{N_{..}} \\
&= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{\left(\frac{N_{i.} N_{.j}}{N_{..}} - N_{ij}\right)^2}{\frac{N_{i.} N_{.j}}{N_{..}}} \\
&= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{N_{..}^2 \left(\frac{N_{i.}}{N_{..}} \frac{N_{.j}}{N_{..}} - \frac{N_{ij}}{N_{..}}\right)^2}{N_{..} \frac{N_{i.}}{N_{..}} \frac{N_{.j}}{N_{..}}} \quad = \quad N_{..} \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{(p_{i.}\, p_{.j} - p_{ij})^2}{p_{i.}\, p_{.j}} \\
&= \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{\frac{1}{N_{..}^2} (N_{i.} N_{.j} - N_{..} N_{ij})^2}{N_{..} \frac{N_{i.}}{N_{..}} \frac{N_{.j}}{N_{..}}} \quad = \quad \sum_{i=1}^{n_C} \sum_{j=1}^{n_A} \frac{(N_{i.}\, N_{.j} - N_{..} N_{ij})^2}{N_{..} N_{i.}\, N_{.j}}.
\end{aligned}$$

This measure is often normalized by dividing it by the size $N_{..} = |\mathcal{X}|$ of the dataset to remove the dependence on the number of sample cases.

For association rule evaluation, C refers the consequent and A to the antecedent of the rule. Both have two values, which we denote by c_0, c_1 and a_0, a_1, respectively. c_0 means that the consequent of the rule is not satisfied, c_1 that it is satisfied; likewise for A. Then we have to compute the χ^2 measure from the 2×2 contingency table

	a_0	a_1	
c_0	N_{00}	N_{01}	$N_{0.}$
c_1	N_{10}	N_{11}	$N_{1.}$
	$N_{.0}$	$N_{.1}$	$N_{..}$

or the estimated probability table

	a_0	a_1	
c_0	p_{00}	p_{01}	$p_{0.}$
c_1	p_{10}	p_{11}	$p_{1.}$
	$p_{.0}$	$p_{.1}$	1

That is, we have

$$\begin{aligned}
\frac{\chi^2(C, A)}{N_{..}} &= \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(p_{i.}\, p_{.j} - p_{ij})^2}{p_{i.}\, p_{.j}}. \\
&= \frac{(p_{0.}\, p_{.0} - p_{00})^2}{p_{0.}\, p_{.0}} + \frac{(p_{0.}\, p_{.1} - p_{01})^2}{p_{0.}\, p_{.1}} + \frac{(p_{1.}\, p_{.0} - p_{10})^2}{p_{1.}\, p_{.0}} + \frac{(p_{1.}\, p_{.1} - p_{11})^2}{p_{1.}\, p_{.1}}
\end{aligned}$$

Now we can exploit

$$p_{00} + p_{01} = p_{0.}, \quad p_{10} + p_{10} = p_{1.}, \quad p_{00} + p_{10} = p_{.0}, \quad p_{01} + p_{11} = p_{.1}, \quad p_{0.} + p_{1.} = 1, \quad p_{.0} + p_{.1} = 1,$$

which leads to

$$\begin{aligned}
p_{0.}\, p_{.0} - p_{00} &= (1 - p_{1.})(1 - p_{.1}) - (1 - p_{1.} - p_{.1} + p_{11}) &= p_{1.}\, p_{.1} - p_{11}, \\
p_{0.}\, p_{.1} - p_{01} &= (1 - p_{1.})p_{.1} - (p_{.1} - p_{11}) &= p_{11} - p_{1.}\, p_{.1}, \\
p_{1.}\, p_{.0} - p_{10} &= p_{1.}(1 - p_{.1}) - (p_{1.} - p_{11}) &= p_{11} - p_{1.}\, p_{.1}.
\end{aligned}$$

Therefore it is

$$\begin{aligned}
\frac{\chi^2(C, A)}{N_{..}} &= \frac{(p_{1.}\, p_{.1} - p_{11})^2}{(1 - p_{1.})(1 - p_{.1})} + \frac{(p_{1.}\, p_{.1} - p_{11})^2}{(1 - p_{1.})\, p_{.1}} + \frac{(p_{1.}\, p_{.1} - p_{11})^2}{p_{1.}(1 - p_{.1})} + \frac{(p_{1.}\, p_{.1} - p_{11})^2}{p_{1.}\, p_{.1}} \\
&= \frac{(p_{1.}\, p_{.1} - p_{11})^2 (p_{1.}\, p_{.1} + p_{1.}(1 - p_{.1}) + (1 - p_{1.})p_{.1} + (1 - p_{1.})(1 - p_{.1}))}{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})} \\
&= \frac{(p_{1.}\, p_{.1} - p_{11})^2 (p_{1.}\, p_{.1} + p_{1.} - p_{1.}\, p_{.1} + p_{.1} - p_{1.}\, p_{.1} + 1 - p_{1.} - p_{.1} + p_{1.}\, p_{.1})}{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})} \\
&= \frac{(p_{1.}\, p_{.1} - p_{11})^2}{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})}.
\end{aligned}$$

In the program, $p_{1.}$ (argument `head`), $p_{.1}$ (argument `body`) and $p_{1|1} = \frac{p_{11}}{p_{.1}}$ (argument `post`, rule confidence) are passed to the routine that computes the measure, so the actual computation is

$$\frac{\chi^2(C, A)}{N_{..}} = \frac{(p_{1.}\, p_{.1} - p_{1|1}\, p_{.1})^2}{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})}. = \frac{((p_{1.} - p_{1|1})p_{.1})^2}{p_{1.}(1 - p_{1.})p_{.1}(1 - p_{.1})}.$$

In an analogous way the measure can also be computed from the absolute frequencies N_{ij}, $N_{i.}$, $N_{.j}$ and $N_{..}$, namely as

$$\frac{\chi^2(C, A)}{N_{..}} = \frac{(N_{1.} N_{.1} - N_{..} N_{11})^2}{N_{1.}(N_{..} - N_{1.})N_{.1}(N_{..} - N_{.1})}.$$