Supplementary Material for
Discretization and Feature Selection Based on Bias Corrected Mutual Information Considering High-Order Dependencies

Puloma Roy\(^1\), Sadia Sharmin\(^2\), Amin Ahsan Ali\(^3\), and Mohammad Shoyaib\(^1\)

\(^1\) Institute of Information Technology, University of Dhaka, Bangladesh
pulomaa92@gmail.com, shoyaib@du.ac.bd
\(^2\) Dept. of CSE, Islamic University of Technology, Bangladesh
sharmin@iut-dhaka.edu
\(^3\) Dept. of CSE, Independent University of Bangladesh, Bangladesh
aminali@iub.edu.bd

Proof of Theorems

**Theorem 1.** Bias is \(\frac{(M-1)(J-1)I}{2N \ln 2}\) for Interaction \(I(f_m; f_j \mid f_i)\) among the features \(f_m\) and \(f_j\) given feature \(f_i\), where \(I, J\) and \(M\) are the number of intervals in feature \(f_i\), \(f_j\) and \(f_m\) respectively.

**Proof.** Conditional mutual information between \(X, Y\) given \(Z\) can be represented as Eq. 1

\[ I(X; Y \mid Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z) \] (1)

For continuous distribution \(H(X, Y, Z)\) can be defined using Eq. 2

\[ H(X, Y, Z) = - \int_\infty^{-\infty} \int_\infty^{-\infty} f(x, y, z) \log f(x, y, z) dx dy dz \] (2)

Let us divide the xyz-space into \((M \times J \times I)\) equally sized \((\Delta x \times \Delta y \times \Delta z)\) cells with coordinates \((m, j, i)\). The number of samples observed in cell \((m, j, i)\) is \(n_{mji}\) and the total number of samples is \(N\). Then, the estimator function of entropy \(H(x, y, z)\) will be

\[ \hat{H}(X, Y, Z) = -\frac{1}{\ln 2} \sum_{m,j,i} \left( \frac{n_{mji}}{N} \ln \frac{n_{mji}}{N} \right) + \log(\Delta x \Delta y \Delta z) \] (3)

If the samples are independent, the stochastic variables \(n_{mji}\) are multinomially distributed. Then, the expectations \(E\{n_{mji}\}\) and variances \(VAR\{n_{mji}\}\) are defined as

\[ E\{n_{mji}\} = \bar{n}_{mji} = N \rho_{mji} \] (4)
\[ \text{VAR}\{n_{mji}\} = Np_{mji}(1 - p_{mji}) \]  

Let us define \( p_1 = \frac{n_{mji}}{N} \) and \( q_1 = \frac{n_{mji}}{N} \) and assume \( f(p_1) = \frac{n_{mji}}{N} \ln \frac{n_{mji}}{N} = p_1 \ln p_1 \) as a function of \( p_1 \). Substituting these value in 3 and expanding \( f(p_1) \) through Taylor series at \( p_1 = q_1 \), we find

\[ f(p_1) = f(q_1) + f^{(1)}(q_1)(p_1 - q_1) + \frac{f^{(2)}(q_1)}{2!}(p_1 - q_1)^2 + \cdots \]

\[ \text{[} f^{(r)}(p_1) \text{ is the } r^{th} \text{ derivative of } f(p_1) \text{]} \]

\[ = q_1 \ln q_1 + (1 + \ln q_1)(p_1 - q_1) + \frac{(p_1 - q_1)^2}{2q_1} \]  

(6)

Applying the value of \( p_1 \) and \( q_1 \) into Eq. 6, Eq. 3 can be written as

\[ \hat{H}(X,Y,Z) = \frac{1}{\ln 2} \sum_{m,j,i} (-\frac{n_{mji}}{N} \ln \frac{n_{mji}}{N} - (\frac{1}{N} + \frac{1}{N} \ln \frac{n_{mji}}{N})(n_{mji} - n_{mji}) - \frac{(n_{mji} - n_{mji})^2}{2Nn_{mji}} + R_{3mji}^3(n_{mji})) + \log(\Delta x \Delta y \Delta z) \]  

(7)

In Eq. 7, \( R_{3mji}^3(n_{mji}) \) is the higher order term of the Taylor expansion. Replacing the formal parameters \( n_{mji} \) by the stochastic variables \( \bar{n}_{mji} \), we take the expectations of \( \hat{H}(X,Y,Z) \) assuming independent samples \( \bar{n}_{mji} \) is multinomially distributed.

\[ E\{\hat{H}(X,Y,Z)\} = \frac{1}{\ln 2} \sum_{m,j,i} (-\frac{\bar{n}_{mji}}{N} \ln \frac{\bar{n}_{mji}}{N} - (\frac{1}{N} + \frac{1}{N} \ln \frac{\bar{n}_{mji}}{N})E\{(\bar{n}_{mji} - n_{mji})\} - \frac{E\{(n_{mji} - n_{mji})^2\}}{2N\bar{n}_{mji}} + E\{R_{3mji}^3(\bar{n}_{mji})\}) + \log(\Delta x \Delta y \Delta z) \]  

(8)

In Eq. 8, the 2nd term vanishes and the 3rd term can be written in the following form using the value of expectations and variance from Eq. 4 and 5

\[ \frac{1}{\ln 2} \sum_{m} \sum_{j} \sum_{i} E\{(n_{mji} - n_{mji})^2\} = \frac{1}{\ln 2} \sum_{m} \sum_{j} \sum_{i} Np_{mji}(1 - p_{mji}) \]  

\[ = \frac{1}{\ln 2} \sum_{m} \sum_{j} \sum_{i} \frac{(1 - p_{mji})}{2N} \]  

\[ = \frac{1}{2N\ln 2} \left( \sum_{m} \sum_{j} \sum_{i} 1 - \sum_{m} \sum_{j} \sum_{i} p_{mji} \right) \]  

\[ = \frac{1}{2N\ln 2} (MJI - 1) \]  

(9)
Theorem 2.

Proof. Conditional mutual information between two independent random variables $X$ and $Y$ for a given $Z$ can be defined as

$$I(X; Y | Z) = \frac{1}{\ln 2} \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \ln \frac{p(x, y, z)}{p(x | z)p(y | z)p(z)}$$  \hspace{1cm} (13)

Here, $p(x, y, z)$ is the joint probability and $p(x | z), p(y | z)$ are the conditional probability. Let $q(x, y, z) \equiv p(x | z)p(y | z)p(z)$.

$$I(X; Y | Z) = \frac{1}{\ln 2} \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \ln \frac{p(x, y, z)}{q(x, y, z)}$$  \hspace{1cm} (14)

For simplicity, assume $p_1 = p(x, y, z), q_1 = q(x, y, z)$ and $f(p_1) = p_1 \ln \frac{p_1}{q_1}$ as a function of $p_1$. Expanding $f(p_1)$ into a Taylor series at $p_1 = q_1$ yields

$$f(p_1) = f(q_1) + f^{(1)}(q_1)(p_1 - q_1) + \frac{f^{(2)}(q_1)}{2!}(p_1 - q_1)^2 + \cdots$$  \hspace{1cm} (15)
where $f'(p_1)$ is the $r^{th}$ derivative of $f(p_1)$. The first term vanishes, the second term is $(p_1 - q_1)$ and the third term is $\frac{(p_1 - q_1)^2}{2q_1}$. Ignoring the higher order term, we can write Eq. 14 as

$$I(X; Y \mid Z) \approx \frac{1}{\ln 2} \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} f(p_1)$$

$$\approx \frac{1}{\ln 2} \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} ((p_1 - q_1) + \frac{(p_1 - q_1)^2}{2q_1})$$

$$\approx \frac{1}{\ln 2} \left( \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) - \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} q(x, y, z) \right.$$  

$$\left. + \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} \frac{(p(x, y, z) - q(x, y, z))^2}{2q(x, y, z)} \right)$$

$$\approx \frac{1}{\ln 2} \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} \frac{(p(x, y, z) - q(x, y, z))^2}{2q(x, y, z)}$$  \hspace{1cm} (16)

As $\sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} q(x, y, z) = 1$

Replacing $p(x, y, z) = \frac{n_{mji}}{N}$, $p(x \mid z) = \frac{p(x, z)}{p(z)} = \frac{n_{m.j}}{n_{.i}}$, $p(y \mid z) = \frac{p(y, z)}{p(z)} = \frac{n_{m.j}}{n_{.i}}$, $p(z) = \frac{n_{..i}}{N}$, Eq. 16 can be written as

$$I(X; Y \mid Z) = \frac{1}{\ln 2} \sum_{m} \sum_{j} \sum_{i} \frac{n_{mji}}{N} \ln \frac{n_{mji}}{n_{m.j}} \frac{n_{m.j}}{n_{..i}} \frac{n_{..i}}{N}$$ \hspace{1cm} (17)

At this point, this expression of conditional mutual information is related to a standard test of independence in statistics, the $\chi^2$ test. The $\chi^2$ test variable is defined as

$$\chi^2 = \sum_{m=1}^{X} \sum_{j=1}^{Y} \sum_{i=1}^{Z} \frac{(n_{mji} - \frac{n_{..i}n_{m.j}}{n_{.i}})^2}{\frac{n_{..i}n_{m.j}}{n_{.i}}}$$ \hspace{1cm} (18)

with $(X - 1)(Y - 1)Z$ degrees of freedom if $X$, $Y$ and $Z$ are statistically independent [1]. Hence, $I(f_m; f_j \mid f_i)$ follows $\chi^2$ distribution.

As Interaction follows $\chi^2$ distribution, its critical value can be calculated using Eq. 19

$$\chi^2_{(ij)} = 2N \ln(2) * I(f_m; f_j \mid f_i)$$ \hspace{1cm} (19)

References
1. Myers, D.E.: Some aspects of multivariate analysis. In: Quantitative Analysis of Mineral and Energy Resources, pp. 669–687. Springer (1988)
2. Sharmin, S., Shoyaib, M., Ali, A.A., Khan, M.A.H., Chae, O.: Simultaneous feature selection and discretization based on mutual information. Pattern Recognition 91, 162–174 (2019)