Nominal Association Measures and Feature Selection for Categorical Data

Xiaogang (Steven) Wang

Dept. of Mathematics and Statistics,
York University

Joint Work with Wenxue Huang and Yong Shi

Dec.12, 2011

Outline

1. Introduction
2. General Nominal Association Measures
3. Simulation and Data Analysis
4. Conclusion and Future Work

Nominal Associations

- We are interested in association measures for nominal data.
- The response variables have more than two categories.
- Goodman and Kruskal have argued that many nominal association measures are based on χ^2 statistic for testing independence.

Goodman and Kruskal's τ

Goodman and Kruskal (1954) proposed the following measure:

$$\tau^{Y|X} = \frac{\sum_{i=1}^{n_Y} \sum_{j=1}^{n_X} p(Y = i; X = j)^2 / p(X = j) - \sum_{i=1}^{n_Y} p(Y = i)^2}{1 - \sum_{i=1}^{n_Y} p(Y = i)^2}, \quad (1)$$

where n_X and n_Y represent the number of classes/categories for X and Y respectively.

Goodman and Kruskal's τ

The principle can be stated as

$$r(Y|X) = [V(Y) - V(Y|X)]/V(Y) \tag{2}$$

where V symbolizes certain measure of uncertainty.

Goodman and Kruskal's τ can be derived
by using Gini concentration index defined as

$$V_G(X) := \sum_{i=1}^{n_X} p(X = i)(1 - p(X = i)). \tag{3}$$

Association Vector and General Association Degree

The *association vector*

$$\Theta^{Y|X} := (\theta^{(Y=1)|X}, \theta^{(Y=2)|X}, \ldots, \theta^{(Y=n_Y)|X}) \quad (4)$$

is given by

$$\theta^{(Y=s)|X} := \frac{E[p(Y=s|X)^2] - p(Y=s)^2}{p(Y=s)\,(1-p(Y=s))}, \qquad s = 1, 2, \ldots, n_Y. \quad (5)$$

The global association degree is defined as

$$\tau_{\boldsymbol{\alpha}}^{Y|X} = \sum_{s=1}^{n_Y} \alpha_s\, \theta^{Y=s|X}. \quad (6)$$

with $\sum_s \alpha_s = 1$ and $\alpha_s \geq 0$ for all $s = 1, 2, \ldots, n_Y$,

Theoretical Properties

THEOREM 1. Assume α is a regular weight vector.

1. $0 \leq \theta^{(Y=s)|X} \leq 1$ and $0 \leq \tau_{\boldsymbol{\alpha}}^{Y|X} \leq 1$;
2. $\tau_{\boldsymbol{\alpha}}^{Y|X} = 0 \iff Y$ and X are independent.
3. $\tau_{\boldsymbol{\alpha}}^{Y|X} = 1 \iff Y$ is completely determined by X;
4. If the weight vector is assigned as in the following:

$$\boldsymbol{\alpha}^P = \frac{1}{V_G(Y)}(p(Y=1)-p(Y=1)^2, \ldots, p(Y=n_Y)-p(Y=n_Y)^2); \tag{7}$$

where $V_G(Y) = \sum_s p(Y=s)(1-p(Y=s))$, then

$$\tau^{Y|X} = \tau_{\boldsymbol{\alpha}^P}^{Y|X}. \tag{8}$$

Association Matrix

The association matrix is given by

$$\gamma(Y|X) := (\gamma^{st}(Y|X)), \tag{9}$$

where

$$\gamma^{st}(Y|X) := \frac{E[p(Y=s|X)\, p(Y=t|X)]}{p(Y=s)},$$

where $s, t = 1, 2, \cdots, n_Y$.

We have the following properties:

(1) $\gamma(Y|X)$ is a row stochastic matrix;

(2) $\theta^{(Y=s)|X}$ is the normalization of $\gamma^{ss}(Y|X)$;

Hierarchical Equivalence Structure

1. *E-1 equivalent*, if $\tau^{X_1|X_2} = \tau^{X_2|X_1} = \tau^{Y|X_1} = 1$;
2. *E-2 equivalent*, if $\tau^{Y|X_1} = 1 = \tau^{Y|X_2}$;
3. *E-3 equivalent*, if $\gamma(Y|X_1) = \gamma(Y|X_2)$;
4. E-4 equivalent, if $\Theta^{Y|X_1} = \Theta^{Y|X_2}$;
5. E-5 equivalent with respect to a weight vector α, if $\tau_\alpha^{Y|X_1} = \tau_\alpha^{Y|X_2}$.

THEOREM 2. If X_1 and X_2 are E-i equivalent (with respect to Y), then they are E-(i+1) equivalent (with respect to Y), for $i = 1, 2, 3, 4$.

Structural Base

Definition

A subset

$$\{V_{i_1}, V_{i_2}, \ldots, V_{i_k}\} \subseteq \mathrm{V}$$

is called an α-structural base for S for a given regular weight vector α if

αB1. for each $V_i \in \mathrm{V}$, $\tau_\alpha^{V_i|(V_{i_1}, V_{i_2}, \ldots, V_{i_k})} = 1$;

αB2. for any $V \in \{V_{i_1}, \ldots, V_{i_k}\}$, $\tau_\alpha^{V|(\{V_{i_1}, \ldots, V_{i_k}\} \setminus \{V\})} < 1$.

THEOREM 3. Let S be a data set with independent variables $X_1, \ldots, X_n$ and a dependent variable Y, and α a weight vector. Then there exists an α-association base.

ALGORITHM

1. Compute all pairwise associations and find the best one.
2. Compute the augmented associations by adding one more variables.
3. Find the best combination from Step 2.
4. Repeat Step 2 -3 until a stopping rule is reached.

The computational complexity for computing each τ_α is approximately $O(n\ m)$. Bootstrapped confidence interval might be needed when the sample size is moderate.

SIMULATION

The joint distribution of $(Y, X1, X2)$ is given by the following:

(X_1, X_2)	$P(X1; X2)$	$P(Y=0\|X_1, X_2)$	$P(Y=1\|X_1; X_2)$	$P(Y=2\|X_1; X_2)$
(0, 0)	9/16	95%	5%	0%
(0, 1)	3/16	30%	70%	0%
(1, 0)	3/16	50%	50%	0%
(1, 1)	1/16	0%	5%	95%

We also generate redundant variables $X3$, $X4$ and $X5$.
$P(X_3 = 1|X_1 = 1) = P(X_4 = 1|X_2 = 1) = 0.90$.
$X_5 = I(X_2 = 1) * I(X_3 = 1) * Z$, where $P(Z = 1) = 0.8$.
Note that $P_Y = (0.6875, 0.2531, 0.0594)$..

Weighting Schemes

For comparison purposes, we consider three weighting schemes:

$$\alpha_k^{GK} = \frac{p(Y=k)(1-p(Y=k))}{\sum_{j=1}^{n_Y} p(Y=j)(1-p(Y=j))}; \tag{10}$$

$$\alpha_k^{EW} = 1/n_Y; \tag{11}$$

$$\alpha_k^{IPW} = 1/p(Y=k)/[\sum_{j=1}^{n_Y} 1/p(Y=j)]; \tag{12}$$

assuming that $p(Y=k) > 0$ for any k.

SIMULATION RESULTS

	X_1	X_2	X_3	X_4	X_5
τ(GK)	0.2382	0.1010	0.2060	0.0878	0.1511
τ(EW)	0.2221	0.1206	0.1923	0.1050	0.2943
τ(IPW)	0.1900	0.1597	0.1648	0.1393	0.5806

	$X_1 + X_4$	$X_2 + X_3 + X_5$	$X_1 + X_4 + X_5$	$X_1 + X_2$	ALL
τ(GK)	0.4627	0.4669	0.4823	0.5018	0.5018
τ(EW)	0.5570	0.5731	0.5884	0.6078	0.6078
τ(IPW)	0.7372	0.7940	0.8004	0.8198	0.8199

DATA ANALYSIS

Data Set: n=24000; m=120.

τ(GK)	0.0825	0.0760	0.0719	0.0680	0.0540
Variables	V34	V2	V48	V8	V12
τ(EW)	0.0911	0.0835	0.0809	0.0787	0.0638
Variables	V34	V2	V8	V48	V12
τ(IPW)	0.1075	0.0888	0.0773	0.0682	0.0605
Variables	V8	V34	V12	V111	V2

Table: Measure of associations using different weighting schemes for significant variables and some of their combinations.

DATA ANALYSIS

Data Set: n=24000; m=120.

τ(GK)	0.1195	0.1136	0.1076	0.1021	0.1013
V34	+V2	+V48	+V8	+V7	+V5
τ(EW)	0.1333	0.1268	0.1268	0.1128	0.1107
V34	+V2	+V8	+V48	+V7	+V5
τ(IPW)	0.1945	0.1779	0.1598	0.1589	0.1569
V8	+V1	+V22	+V34	+V2	+V48

Table: Top five measure of associations based on the combination of two variable. The best single variables are three weighting schemes are V34, V34, and V2, V8 respectively.

DATA ANALYSIS

Data Set: n=24000; m=120.

τ(GK) V34+V2	0.1454 +V8	0.1397 +V102	0.1393 +V5	0.1380 +V12	0.1368 +V111
τ(EW) V34+V2	0.1709 +V8	0.1583 +V12	0.1563 +V111	0.1554 +V102	0.1554 +V5
τ(IPW) V8+V1	0.2576 +V2	0.2527 +V48	0.2379 +V7	0.2333 +V34	0.2321 +V102

Table: Top five measure of associations based on the combinations of triple variables by using V34+V2 and another explanatory variable using different weighting schemes.

Summary

- We introduce a association vector and matrix to measure nominal associations.
- These measures provide both local and global evaluations.
- We also discover the hierarchical equivalence structure.
- The existence of a structural basis is proved.

Future Works

- Extension to ordinal categorical data.
- Bayesian approach when the sample size is small.
- Application to risk management and biological data.

References

Huang, W., Shi, Y and Wang, X. (2011). Nominal Association Measures and Feature Selection for Categorical Data. To be submitted.

Goodman, L.A. (1996). JASA **91**, 408-428.

Goodman, L.A. and Kruskal, W. H. (1954). JASA, **49**, 732-764.

Lloyd, C. J. (1999). Statistical analysis of categorical data.