Aggregation and Association in Cross Tables

Gilbert Ritschard[1] and Nicolas Nicoloyannis[2]

[1] Dept of Econometrics, University of Geneva, CH-1211 Geneva 4, Switzerland
ritschard@themes.unige.ch
[2] Laboratoire ERIC, University of Lyon 2, C.P.11 F-69676 Bron Cedex, France
nicolas.nicoloyannis@univ-lyon2.fr

Abstract. The strength of association between the row and column variables in a cross table varies with the level of aggregation of each variable. In many settings like the simultaneous discretization of two variables, it is useful to determine the aggregation level that maximizes the association. This paper deals with the behavior of association measures with respect to the aggregation of rows and columns and proposes an heuristic algorithm to (quasi-)maximize the association through aggregation.

1 Introduction

This paper is concerned with the maximization of the association in a cross table. In this perspective, it deals with the effect of the aggregation of the row and column variables on association measures and proposes an aggregation algorithm to (quasi-)maximize given association criteria. Indeed, the strength of the association may vary with the level of aggregation. It is well known for instance that aggregating identically distributed rows or columns does not affect the Pearson or Likelihood ratio Chi-squares (see for example [4], p. 450) but increases generally the value of association measures as illustrated for instance by the simulations carried out in [5].

Maximizing the strength of the association is of importance in several fields. For example, when analyzing survey sample data it is common to group items to avoid categories with too few cases. For studying the association between variables, as it is often the case in social sciences, it is, then, essential to understand how the association may vary with the grouping of categories in order to select the grouping that best reflects the association. This issue was discussed, for instance, by Benzécri [1] with respect to the maximization of the Pearson's Chi-square.

A second motivation concerns discretization that is a major issue in supervised learning. In this framework, a joint aggregation of the predictor and the response variable should be more efficient than an independent discretization of each variable. Nevertheless, with the exception of the joint dichotomization process considered by Breiman et al. [2], it seems that optimal solutions for partitioning values exist in the literature only for a single variable (see for instance [8] for a survey). There is thus an obvious need for tools allowing a joint general

optimal discretization of two or more variables. The results presented here for jointly partitioning the row and column values in an unfixed number of classes are a first step in this direction.

The joint optimal partition can be found by scanning all possible groupings. Since the number of these groupings increases exponentially with the number of categories, such a systematic approach is, however, generally untractable. An iterative process is thus introduced for determining the (quasi-)optimal partition by successively aggregating two categories.

The distinction between nominal and ordinal variables is of first importance in this partitioning issue. Indeed, with ordinal variables only an aggregation of adjacent categories makes sense.

It is worth mentioning that the optimization problem considered here differs from that of Benzécri [1], for which Celeux et al. [3] have proposed an algorithm based on clustering techniques. These authors consider only partitions into a fixed number of classes and maximize Pearson's Chi-square. Unlike this framework, our settings allow varying numbers of rows and columns. We have therefore to rely on normalized association measures to compare configurations.

Section 2 illustrates with an example how the aggregation of row and column values may affect the Chi-square statistics and the association measures. The formal framework and the notations are described in Section 3. Section 4 specifies the complexity of the enumeration search of the optimal solution and proposes an heuristic algorithm heuristic. Section 5 summarizes a sensitivity analysis of association measures. Finally, Section 6 proposes some concluding remarks.

2 Example and Intuitive Results

Consider the following cross table between a row variable x and a column variable y

$$M = \begin{array}{c|cccc} x\backslash y & A & B & C & D \\ \hline a & 10 & 10 & 1 & 1 \\ b & 10 & 10 & 1 & 1 \\ c & 1 & 1 & 10 & 10 \\ d & 1 & 1 & 10 & 10 \end{array} .$$

Intuitively, aggregating two identical columns, $\{A, B\}$ or $\{C, D\}$, should increase the association level. The same holds for a grouping of rows $\{a, b\}$ or $\{c, d\}$. On the other hand, grouping categories B and C for example reduces the contrast between column distributions and should therefore reduce the association. Let us illustrate these effects by considering the aggregated tables

$$M_y^+ = \begin{array}{c|ccc} x\backslash y & A & B & \{C, D\} \\ \hline a & 10 & 10 & 2 \\ b & 10 & 10 & 2 \\ c & 1 & 1 & 20 \\ d & 1 & 1 & 20 \end{array} \qquad M_{xy}^+ = \begin{array}{c|ccc} x\backslash y & A & B & \{C, D\} \\ \hline a & 10 & 10 & 2 \\ b & 10 & 10 & 2 \\ \{c, d\} & 2 & 2 & 40 \end{array}$$

and

$$M_y^- = \begin{array}{c|ccc} x\backslash y & A & \{B,C\} & D \\ \hline a & 10 & 11 & 1 \\ b & 10 & 11 & 1 \\ c & 1 & 11 & 10 \\ d & 1 & 11 & 10 \end{array} \qquad M_{xy}^- = \begin{array}{c|ccc} x\backslash y & A & \{B,C\} & D \\ \hline a & 10 & 11 & 1 \\ \{b,c\} & 11 & 22 & 11 \\ d & 1 & 11 & 10 \end{array} .$$

Table 1. Association measures for different groupings

	M	M_y^+	M_{xy}^+	M_y^-	M_{xy}^-
rows	4	4	3	4	3
columns	4	3	3	3	3
df	9	6	4	6	4
Pearson Chi-square	58.91	58.91	58.91	29.45	14.73
Likelihood Ratio	68.38	68.38	68.38	34.19	17.09
Tschuprow's t	0.47	0.52	0.58	0.37	0.29
Cramer's v	0.47	0.58	0.58	0.41	0.29
Goodman-Kruskal $\tau_{y\leftarrow x}$	0.22	0.40	0.40	0.13	0.07
Goodman-Kruskal $\tau_{x\leftarrow y}$	0.22	0.22	0.40	0.11	0.07
Uncertainty Coefficient $u_{y\leftarrow x}$	0.28	0.37	0.37	0.19	0.09
Uncertainty Coefficient $u_{x\leftarrow y}$	0.28	0.28	0.37	0.14	0.09
Goodman-Kruskal γ	0.68	0.77	0.80	0.63	0.57
Kendall's τ_b	0.55	0.60	0.65	0.45	0.37
Somers' $d_{y\leftarrow x}$	0.55	0.55	0.65	0.41	0.37
Somers' $d_{x\leftarrow y}$	0.55	0.65	0.65	0.49	0.37

Table 1 gives the values of a set of nominal (t, v, τ, u) and ordinal (γ, τ_b, d) association measures for cross table M and the four aggregated tables considered. The nominal τ and u and the ordinal d are directional measures. For more details on these measures see fo instance [5]. According to the distributional equivalence property, the Chi-square statistics remain the same for tables M, M_y^+ and M_{xy}^+. However, the association measures increase as expected with the grouping of similar columns and rows. The figures for the aggregated tables M_y^- and M_{xy}^- show that the aggregation of columns or rows very differently distributed reduces both the Chi-squares and the association measures.

3 Notations and Formal Framework

Let x and y be two variables with respectively r and c different states. Crossing variable x with y generates a contingency table $T_{r\times c}$ with r rows and c columns. Let $\theta_{xy} = \theta(T_{r\times c})$ denote a generic association criterion for table $T_{r\times c}$. Let P_x be a partition of the values of x, and P_y a partition of the states of y. Each couple

(P_x, P_y) defines then a contingency table $T(P_x, P_y)$. The optimization problem considered is the maximization of the association among the set of couples of partitions (P_x, P_y)

$$\max_{P_x, P_y} \theta\big(T(P_x, P_y)\big) \tag{1}$$

For ordinal variables, hence for interval or ratio variables, only partitions obtained by aggregating adjacent categories make sense. We consider then the restricted program

$$\begin{cases} \max\limits_{P_x, P_y} \theta\big(T(P_x, P_y)\big) \\ \text{u.c.} \quad P_x \in \mathcal{A}_x \text{ and } P_y \in \mathcal{A}_y \end{cases} \tag{2}$$

where $\mathcal{A}_x$ and $\mathcal{A}_y$ stand for the sets of partitions obtained by grouping adjacent values of x and y. Letting $\mathcal{P}_x$ and $\mathcal{P}_y$ be the unrestricted sets of partitions, we have for $c, r > 2$, $\mathcal{A}_x \subset \mathcal{P}_x$ and $\mathcal{A}_y \subset \mathcal{P}_y$. Finally, note that for ordinal association measures that may take negative values, maximizing the strength of the association requires to take the absolute value of the ordinal association measure as objective function $\theta\big(T(P_x, P_y)\big)$.

4 Complexity of the Optimal Solution

4.1 Complexity of the Enumerative Approach

To find the optimal solution, we have to explore all possible groupings of both the rows and the columns, i.e. the set of couples (P_x, P_y). The number of cases to be checked is given by $\#\mathcal{P}_x \#\mathcal{P}_y$, i.e. the number of row groupings times the number of column groupings.

For a nominal variable, the number of possible groupings is the number $B(c) = \#\mathcal{P}$ of partitions of its c categories. It may be computed iteratively by means of Bell formula

$$B(c) = \sum_{k=0}^{c-1} \binom{c-1}{k} B(k)$$

with $B(0) = 1$. For $c = r$, the number $B(c)B(r)$ of configurations to be explored is then for example respectively 25, 225, $2'704$ and $41'209$ for $c = 3, 4, 5, 6$ and exceeds $13 \cdot 10^9$ for $c = r = 10$.

For ordinal variables, hence for discretization issues, only adjacent groupings are considered. This reduces the number of cases to browse. The number $G(c) = \#\mathcal{A}$ of different groupings of c values is

$$G(c) = \sum_{k=0}^{c-1} \binom{c-1}{k} = 2^{(c-1)} \ .$$

There are thus respectively $G(c)G(r) = 16$, 64, 256, $1'024$ configurations to browse for a square table with $c = r = 3, 4, 5, 6$, and more than a million for $c = r = 10$.

4.2 An Heuristic

Due to the limitation of the enumerative approach, we propose an iterative process that successively aggregates the two row or column categories that most improve the association criteria $\theta(T)$. Such an heuristic may indeed not end up with the optimal solution, but perhaps only with a quasi-optimal solution.

Formally, the configuration (P_x^k, P_y^k) obtained at step k is the solution of

$$\begin{cases} \max\limits_{P_x, P_y} \ \theta\big(T(P_x, P_y)\big) \\ \text{u.c. } P_x = P_x^{(k-1)} \text{ and } P_y \in \mathcal{P}_y^{(k-1)} \\ \qquad \text{or} \\ \qquad P_x \in \mathcal{P}_x^{(k-1)} \text{ and } P_y = P_y^{(k-1)} \end{cases}, \tag{3}$$

where $\mathcal{P}_x^{(k-1)}$ stands for the set of partitions on x resulting from the grouping of two classes of the partition $P_x^{(k-1)}$.

For ordinal variables, $\mathcal{P}_x^{(k-1)}$ and $\mathcal{P}_y^{(k-1)}$ should be replaced by the sets $\mathcal{A}_x^{(k-1)}$ and $\mathcal{A}_y^{(k-1)}$ of partitions resulting from the aggregation of two adjacent elements.

Starting with $T^0 = T_{r\times c}$ the table associated to the finest categories of variables x and y, the algorithm successively determines the tables $T^k, k = 1, 2, \ldots$ corresponding to the partitions solution of (3). The process continues while $\theta(T^k) \geq \theta(T^{(k-1)})$ and is stopped when the best grouping of two categories leads to a reduction of the criteria.

The *quasi-optimal grouping* is the couple (P_x^k, P_y^k) solution of (3) at the step k where

$$\theta\big(T^{(k+1)}\big) - \theta\big(T^k\big) < 0 \quad \text{and} \quad \theta\big(T^k\big) - \theta\big(T^{(k-1)}\big) \geq 0$$

By convention, we set the value of the association criteria $\theta(T)$ to zero for any table with a single row or column. The algorithm then ends up with such a single value table, if and only if all rows (columns), are equivalently distributed.

5 Effect of Aggregating Two Categories

For the heuristic proposed, it is essential to understand how the association criteria behave in response to the aggregation of two categories. We have therefore carried out an analytical sensitivity analysis reported in [7] of which we summarize here the main results.

Chi-square statistics remain constant when the two aggregated categories are equivalently distributed and decrease otherwise.

Chi-square based measures: Tschuprow's t can increase when r or c decreases. Cramer's v may only increase when the aggregation is done on the variable with the smaller number ($\min\{r, c\}$) of categories.

Nominal PRE measures: $\tau_{y\leftarrow x}$ and $u_{y\leftarrow x}$ may only increase with an aggregation on the dependent variable.

Ordinal measures: Their absolute value may increase for an aggregation on any variable. In the case of an aggregation of two equivalently distributed categories $|\gamma|$ and $|\tau_b|$ increase while $|\tau_c|$ increases if $\min\{r, c\}$ decreases and $|d_{y \leftarrow x}|$ increases when the aggregation is done on the independent variable x and remains constant otherwise.

6 Further Developments

This paper is concerned with the issue of finding the partitions of the row and column categories that maximizes the association. The results presented, on the complexity of the solution and on the sensitivity of the association criteria toward aggregation, are only preliminary materials. A lot remains to be done, especially concerning the properties and implementation of the algorithm sketched in Section 4.2. Let us just mention two important aspects. First, we have to empirically assess the efficiency of the heuristic. We are presently building simulation designs to check how the quasi-optimal solution provided by the algorithm may differ from the true global optimal solution. From the results of Section 4.1, the comparison with the true solution is only possible for reasonably sized starting tables, say tables with six rows and six columns. Secondly, it is worth to take account of the higher reliability of large figures. The algorithm will therefore be extended by implementing Laplace's estimates of the probabilities to increase the robustness of the solution.

References

1. J.-P. Benzécri. *Analyse des données. Tome 2: Analyse des correspondances*. Dunod, Paris, 1973.
2. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Chapman and Hall, New York, 1993.
3. G. Celeux, E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy. *Classification automatique des données*. Informatique. Dunod, Paris, 1988.
4. J. D. Jobson. *Applied Multivariate Data Analysis. Vol.II Categorical and Multivariate Methods*. Springer-Verlag, New York, 1992.
5. M. Olszak and G. Ritschard. The behaviour of nominal and ordinal partial association measures. *The Statistician*, 44(2):195–212, 1995.
6. R. Rakotomalala and D. A. Zighed. Mesures PRE dans les graphes d'induction: une approche statistique de l'arbitrage généralité-précision. In G. Ritschard, A. Berchtold, F. Duc, and D. A. Zighed, editors, *Apprentissage: des principes naturels aux méthodes artificielles*, pages 37–60. Hermes Science Publications, Paris, 1998.
7. G. Ritschard, D. A. Zighed, and N. Nicoloyannis. Optimal grouping in cross tables. Technical report, Department of Econometrics, University of Geneva, 2000.
8. D. A. Zighed, S. Rabaseda, R. Rakotomalala, and F. Feschet. Discretization methods in supervised learning. *Encyclopedia of Computer Science and Technology*, 40(25):35–50, 1999.
9. D. A. Zighed and R. Rakotomalala. *Graphes d'induction: apprentissage et data mining*. Hermes Science Publications, Paris, 2000.