

Bi Dimensional Partitioning with Limited Loss of Information

Partizioni bi-dimensionali con perdita limitata di informazione

Zighed Djamel A.

Laboratoire ERIC
Université Lyon 2
zighed@univ-lyon2.fr

Ritschard Gilbert

Département d'économétrie
Université de Genève
ritschard@themes.unige.ch

Erray Walid

Laboratoire ERIC
Université Lyon 2
walid.erray@univ-lyon2.fr

Riassunto: L'aggregazione delle righe e colonne di una tabella a doppia entrata ha un costo in termini di informazione perduta. Con riferimento all'informazione che una variabile fornisce rispetto all'altra, viene affrontato il problema dell'aggregazione con l'obiettivo di massimizzare una misura d'associazione tra la variabile di riga e quella di colonna (ad es. gli indici t di Tschuprow, τ di Goodman e Kruskal, τ_b di Kendall). Viene presentata una procedura euristica introdotta da Ritschard e Nikoloyannis (2000), della quale viene discussa la complessità e l'attendibilità. Tale euristica riduce drasticamente la complessità propria della procedura esaustiva consistente nel vaglio di tutti i possibili gruppi. L'attendibilità viene valutata mediante una serie di studi di simulazione. I risultati mostrano che la riduzione di complessità ha un costo relativamente alto in termini di percentuale di ottimi mancanti. Tuttavia si evidenzia che le soluzioni mancate sono molto vicine agli altri ottimi globali. .

Keywords: Optimal Crosstable Reduction, Association, Bi Dimensional Categorization.

1. Introduction

When investigating the relationship between the row and column variables in a cross table, too much empty cells or cells with small frequencies may lead to unreliable results. The usual remedy consists in aggregating categories. This should be done carefully, however, to avoid as much as possible loss of information. If we focus on the information that one variable gives on the other one, i.e. the association between them, the objective becomes the determination of the aggregation of the row and column variables that maximizes their association. Indeed, the strength of the association varies with the level of aggregation. It is well known for instance that if merging identically distributed rows or columns does not affect the Pearson or Likelihood ratio Chi-squares (see for example Jobson, 1992, p. 450), it increases generally the value of association measures as illustrated for instance by the simulations reported in Olszak and Ritschard (1995).

Grouping values of variables is the concern of categorization and discretization. In this framework, the univariate approach that handles each variable independently has been abundantly studied since Fisher (1958). The main distinction is here between contextual and non contextual approaches. The former seeks the partitioning that maximizes a contextual criteria like the predictive power of a classifier. See for instance Zighed et al. (1999) for a survey.

The effect of the groupings on one variable upon the association with a second variable depends obviously on those made on the latter. It is then natural to consider a multidimensional approach that proceeds simultaneously to the aggregation of both variables.

The literature about this multidimensional case is less rich. We can mention the related work by Fisher (1962) about the optimal grouping of the unknowns and equations of predictive economic models. The simultaneous partitioning of the cases and the variables in a data matrix has been studied among others by Anderberg (1973), Bock (1979) and Govaert (1995). Govaert (1984, 1995) investigates the special case of binary tables. In the framework of contingency tables considered in this paper, the optimal partitioning problem has been studied from different points of view. Benzécri (1973) is interested in the partition in a fixed number of groups that maximizes the Pearson Chi-square. Celeux et al. (1988) propose a solution to this problem in the form of an iterative heuristic that clusters alternatively the rows and the columns. Gilula and Krieger (1983) study how the Pearson Chi-square behaves when the table is reduced by aggregation. Hirotsu (1983) and Greenacre (1988) are interested in finding the most homogeneous tables. Breiman et al. (1993) have studied the joint dichotomization of two variables. Their solution is implemented in their C&RT procedure.

The more general multivariate discretization with more than two variables has received some attention from the data mining community. The more significant contributions seem to be those by Dougherty et al. (1995) who look for the groupings that optimize the predictive accuracy of classifiers, and the non contextual techniques proposed by Monti and Cooper (1999) and Bay (2001).

Our objective is to find both the number of groups and the joint partition of rows and columns of a contingency table that maximizes the association. None of the works cited gives a satisfactory solution to this problem. Some, like those done after Benzécri or those by Breiman et al. assume the number of groups fixed a priori. The others either do not consider the case of contingency tables or consider criteria, homogeneity for example, that are hardly transposable in our setting.

Clearly, the exhaustive scanning of all combinations of partitions of each of the variable is not practicable for large tables. We have thus proposed an heuristic algorithm in Ritschard and Nicoloyannis (2000) that successively looks for the optimal grouping of two row or column categories. The effect of the merging of two categories on the main association measures has been examined analytically in Ritschard et al. (2001). The purpose of this paper is to further investigate the complexity and the efficacy of the algorithm by completing the preliminary analysis presented in Ritschard et al. (2001); Ritschard (2001).

The formal framework and the notations are defined in Section 2. The heuristic is described in Section 3. Section 4 discusses the complexity of the heuristic and compares it with that of the enumeration search of the optimal solution. Section 5 exhibits the outcomes of a series of simulation studies carried out to assess the reliability of the heuristic. Finally, Section 6 proposes some concluding remarks.

2. 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 $T_{r \times c}$. This criterion θ_{xy} may thus be the Cramer v or Kendall τ_b for table $T_{r \times c}$.

Let P_x be a partition of the values of the row variable 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 then the maximization of the association θ_{xy} among the set of couples (P_x, P_y) of partitions

$$\max_{P_x, P_y} \theta(T(P_x, P_y)) . \quad (1)$$

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

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

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 ordinal association measures may take negative values. Then, for maximizing the strength of the association, the objective function $\theta(T(P_x, P_y))$ should be the absolute value of the ordinal association measure.

3. The Heuristic

The heuristic, first introduced in Ritschard and Nicoloyannis (2000), is an iterative process that successively merges the two rows or columns that most improve the association criteria $\theta(T)$. Formally, the configuration (P_x^k, P_y^k) obtained at step k is the solution of

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

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 $T^k, k = 1, 2, \dots$ 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

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

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.

4. Complexity

This section compares the complexity of the heuristic to that of the exhaustive exploration of all possible couples (P_x, P_y) of row and column mergings.

For the exhaustive scanning, the number of cases to explore is given by $\#P_x \#P_y$, i.e. the number of row partitions times the number of column partitions.

Consider first the case of a nominal variable. The number $B(c) = \#P$ of possible partitions of its c categories can be computed iteratively by means of Bell formula

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

with $B(0) = 1$. Hence the number of cases to browse is $B(r)B(c)$ in the nominal case.

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

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

Thus, the number of cases to explore is $G(r)G(c)$ in the ordinal case.

For the heuristic, we only give an upper bound for the number of couples (P_x, P_y) we

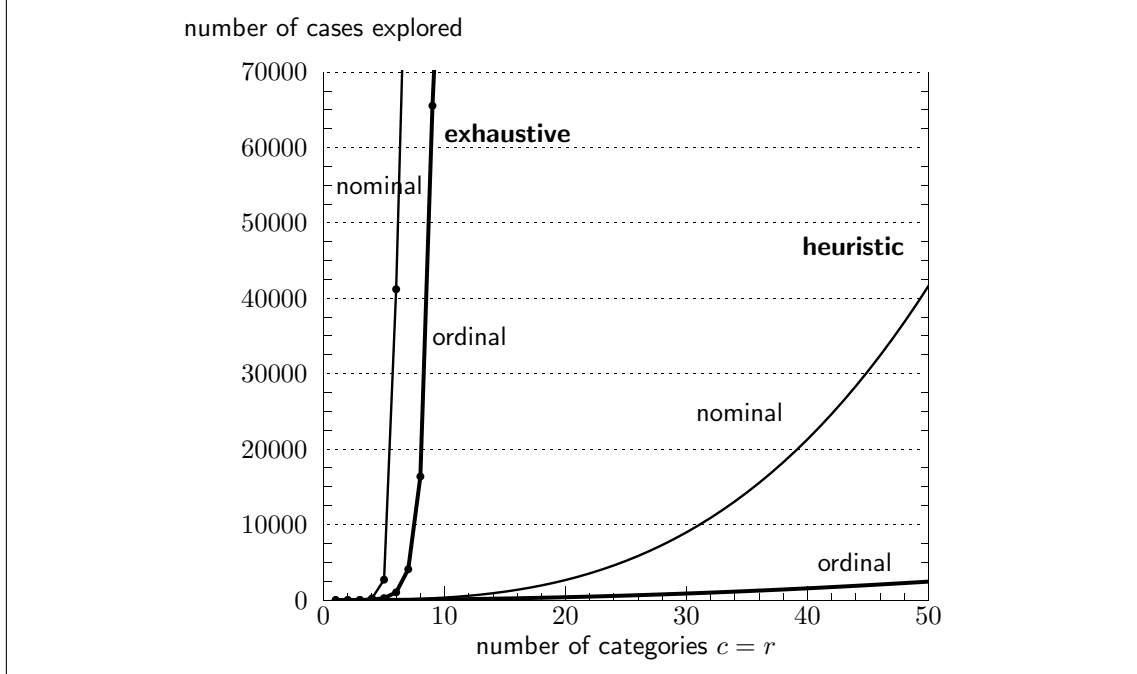


Figure 1: Complexity versus size of the square table (for heuristic, values reported are upper bounds.)

have to scan. This upper bound is

$$\frac{r(r^2 - 1) + c(c^2 - 1)}{6} \quad \text{nominal case}$$

$$\frac{r(r - 1) + c(c - 1)}{2} \quad \text{ordinal case .}$$

Figure 1 and Tables 1 and 2 show how the relative efficacy of the heuristic increases with the number of initial categories. The values reported concern square tables. It is worth mentioning that the seemingly exponential increase in the number of cases reported for the heuristic concerns the upper bound. Practically, the effective number of cases browsed will be much lower.

Table 1: Number of configurations explored, nominal case

$c = r$	exhaustive	heuristic	% gain
2	4	2	50.00%
3	25	8	68.00%
4	225	20	91.11%
5	2704	40	98.52%
6	41209	70	99.83%
7	769129	112	99.99%
8	17139600	168	100.00%
9	447195609	240	100.00%
10	$1.345 \cdot 10^{10}$	330	100.00%
20	$2.675 \cdot 10^{27}$	2660	100.00%
50	$3.449 \cdot 10^{94}$	41650	100.00%
100	$2.264 \cdot 10^{231}$	333300	100.00%

Table 2: Number of configurations explored, ordinal case

$c = r$	exhaustive	heuristic	% gain
2	4	2	50.00%
3	16	6	62.50%
4	64	12	81.25%
5	256	20	92.19%
6	1024	30	97.07%
7	4096	42	98.97%
8	16384	56	99.66%
9	65536	72	99.89%
10	262144	90	99.97%
20	$2.749 \cdot 10^{11}$	380	100.00%
50	$3.169 \cdot 10^{29}$	2450	100.00%
100	$4.017 \cdot 10^{59}$	9900	100.00%

5. Reliability of the heuristic

The purpose of this section is to assess the reliability of the results provided by the heuristic. A series of simulation studies have been run to investigate two aspects: i) the number of global optima missed by the heuristic and ii) how far the quasi-optimum provided by the heuristic is from the global one.

Table 3: Simulations: t of Tschuprow

Tschuprow	nominal			ordinal		
Size	4×4	5×5	6×6	4×4	5×5	6×6
Non zero deviations	39.5%	62.5%	74.5%	23.5%	36%	46.5%
maximum	0.073	0.074	0.077	0.077	0.063	0.108
mean	0.025	0.023	0.028	0.019	0.019	0.012
standard deviation	0.015	0.014	0.016	0.014	0.016	0.015
skewness	0.986	0.979	0.598	1.674	0.972	3.394
With zero deviations						
mean	0.010	0.015	0.021	0.005	0.007	0.006
standard deviation	0.016	0.016	0.018	0.011	0.013	0.012
skewness	1.677	1.062	0.615	3.168	2.211	4.457
Relative deviations						
maximum	0.168	0.198	0.221	0.179	0.194	0.307
mean	0.079	0.077	0.093	0.063	0.066	0.046
Initial mean value	0.260	0.240	0.226	0.263	0.244	0.228
Optima mean value	0.340	0.316	0.303	0.301	0.275	0.250

Table 4: Simulations: τ of Goodman and Kruskal

G&K τ	nominal			ordinal		
Size	4×4	5×5	6×6	4×4	5×5	6×6
Non zero deviations	5%	6.5%	12%	6%	19%	32.5%
maximum	0.013	0.031	0.029	0.076	0.077	0.059
mean	0.007	0.010	0.008	0.025	0.016	0.013
standard deviation	0.004	0.009	0.009	0.021	0.016	0.012
skewness	-0.308	1.004	1.181	1.107	2.361	1.908
With zero deviations						
mean	0.0004	0.0007	0.0010	0.0015	0.003	0.004
standard deviation	0.0018	0.0033	0.004	0.008	0.009	0.009
skewness	5.323	6.471	5.137	6.685	5.040	3.255
Relative deviations						
maximum	0.142	0.296	0.318	0.420	0.518	0.401
mean	0.062	0.091	0.079	0.216	0.168	0.149
Initial mean value	0.074	0.060	0.048	0.073	0.060	0.050
Optima mean value	0.148	0.128	0.113	0.118	0.098	0.084

Several association measures have been examined. We report outcomes for the t of Tschuprow, the τ of Goodman and Kruskal and the τ_b of Kendall. Among the measures

considered¹ the t of Tschuprow has been retained because it provides the worse scores for both the number of missed optima and the deviations from the global optima. The τ of Goodman and Kruskal has been selected as a representative of the asymmetrical PRE (proportion of reduction in error of prediction) measures. Likewise, the τ_b of Kendall has been selected to represent the ordinal measures.

The comparison between quasi and global optima is done for square tables of size 4, 5 and 6. Above 6, the global optimum can no longer be obtained in a reasonable time.

For the t of Tschuprow and the τ of Goodman and Kruskal, we report respectively in Tables 3 and 4 results for the nominal case as well as for the ordinal case. The τ_b of Kendall being an ordinal measure, Table 5 exhibits only figures for the ordinal case.

Table 5: Simulations: τ_b of Kendall

Kendall τ_b	ordinal		
Size	4×4	5×5	6×6
Non zero deviations	19%	24.5%	32%
maximum	0.596	0.597	0.542
mean	0.235	0.182	0.140
standard deviation	0.195	0.190	0.157
skewness	0.076	0.598	0.652
With zero deviations			
mean	0.045	0.045	0.045
standard deviation	0.125	0.123	0.111
skewness	2.775	2.849	2.445
Relative deviations			
maximum	1.954	1.970	1.982
mean	0.355	0.259	0.074
Initial mean value	0.094	0.078	0.064
Optima mean value	0.256	0.236	0.215

For each measure, size and variable type, 200 contingency tables have been randomly generated. Each table was obtained by distributing 10000 cases among its $r \times c$ cells with a random uniform process. This differs from the solution used to generate the results given in Ritschard et al. (2001), which were obtained by distributing the cases with nested conditional uniform distributions: first a random percentage of the cases is attributed to the first row, then a random percentage of the remaining cases is affected to the second row and so on until the last row; the total of each row is then likewise distributed among the columns. The solution retained here generates indeed tables that are closer to the uniform distribution and should therefore exhibit lower association. As will be shown, low association are the less favorable situations for the heuristic. Thus, we can expect the results obtained with this random uniform generating process to provide some upper bounds for the deviations from the global optima.

Tables 3 to 5 exhibit, for each series of tables generated, the proportion of optima missed and characteristic values (maximum, mean value, standard deviation, skewness) of the distribution of the deviations (non zero only and all together.) Relative deviations,

¹Simulations have been run for all the measures discussed in Ritschard et al. (2001), i.e. Tschuprow's t , Cramer's v , Goodman & Kruskal's τ , Theil's u , Goodman and Kruskal's γ , Kendall's τ_b and Somers' d . The latter three are ordinal measures and apply therefore only to ordinal variables.

of which the maximum and the mean value are reported, are the ratios between deviations and global optima. The last two rows give respectively the average of the initial values of the criterion and the mean value of the global optima. In Table 5, these two last figures are means of absolute values of the τ_b 's which may be negative.

Additional insight for the Tschuprow's t and Goodman and Kruskal's τ cases is pro-

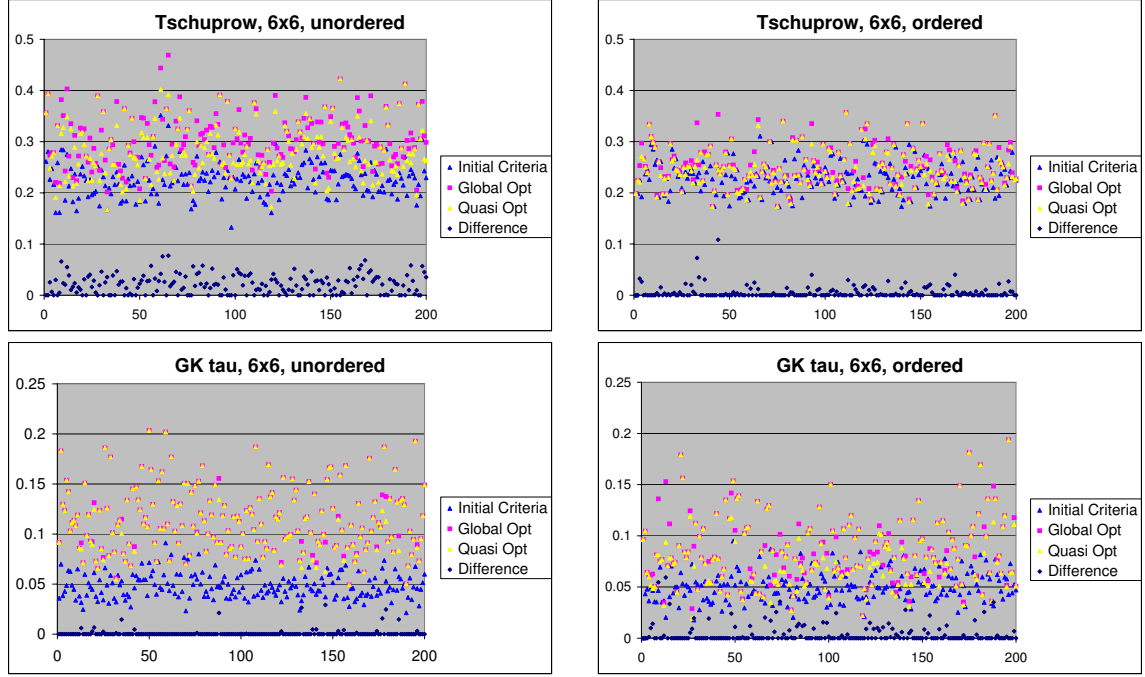


Figure 2: Initial, quasi and global optima

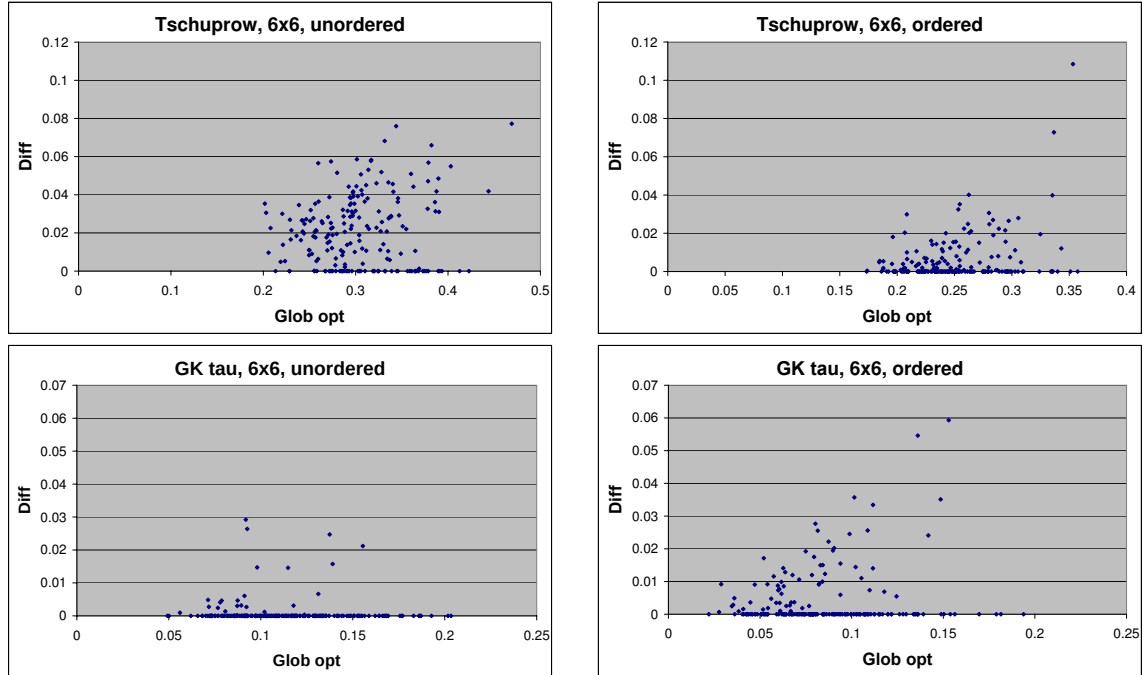


Figure 3: Deviations versus global optima

vided by Figures 2 and 3. Figure 2 shows plots of the 200 initial values, quasi optima and global optima for 6×6 cases. Figure 3 plots the 200 deviations against the global optima.

Looking at Tables 3 to 5, we see that the proportion of optima missed by the heuristic is relatively important and tends to increase with the size of the table. The proportion is somewhat lower for PRE measures (the τ of Goodman and Kruskal). This is probably due to the fact that PRE measures cannot be improved by merging values of the independent variable (see Ritschard et al., 2001), which means that the groupings are in this case almost exclusively made on one (the dependent) variable. Curiously however, the percentages of missed optima are, for PRE measures, larger in the ordinal case than in the nominal one.

This high percentage of missed optima is luckily balanced by the small deviation between the quasi and global optima. The mean value of the non zero deviations is roughly less than half the difference between the initial value of the criterion and the global optimum. In the case of stronger initial associations than those generated here with a uniform random distribution, this ratio becomes largely more favorable, i.e. smaller. The level and dispersion of the non zero deviations seems to remain stable when the size of the table increases. These deviations tend naturally to be larger when the association measures provides larger values. Inversely, the relative deviations take larger values when the association measure tends to zero.

Finally, let us recall that the τ_b of Kendall takes its values in $[-1, 1]$. The deviations may thus exceed the value of the global optimum when the quasi and global optima are of opposite signs. This explains why some maximal relative deviations are greater than one.

Globally, the outcomes of these simulation studies that the cost in terms of reliability of the heuristic remains moderate when compared with the dramatic increase of performance.

6. Further Developments

The issue considered is the finding of the optimal way to merge row and column categories in a crosstable. We focused on the maximization of the association in order to preserve the information that one variable provides on the other one. Other criteria may obviously be considered and should be investigated. For instance, when the data collected in the table are a sample of a larger population, the association computed is an estimate and one should then also care about its standard error. Beside this aspect we are presently working on a strategy to find an optimal aggregation under some constraints. Indeed, following our introductory argument, the primary objective of the reduction of the size of the table is to avoid cells with low frequencies. Therefore, it is worth to be able to maximize for instance the association under a constraint on the minimal cell frequency. Likewise, if the row and column variables serve to control a stratification for estimating some value, it may be of interest to find the aggregation that minimizes the dispersion of the estimator under the constraint of a minimal frequency for the strata. On the algorithmic side, we are working on a top-down hierarchical approach in which, starting from the completely aggregated table we would iteratively split rows or columns.

Finally, let us mention that the heuristic discussed has been implemented in the form of a self standing program for the Windows environment. The software is named OCA for Optimal Crosstable Aggregation. It is menu driven and comes with a help and examples files. It is freely downloadable at <http://mephisto.unige.ch/stats>.

References

- Anderberg, M. (1973). *Cluster Analysis for Application*. New York: Academic Press.
- Bay, S. D. (2001). Multivariate discretization for set mining. *Knowledge and Information Systems* 3(4), 491–512.
- Benzécri, J.-P. (1973). *Analyse des données. Tome 2: Analyse des correspondances*. Paris: Dunod.
- Bock, H. (1979). Simultaneous clustering of objects and variables. In E. Diday et al. (Ed.), *Data Analysis and Informatics*, pp. 187–203.
- Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1993). *Classification And Regression Trees*. New York: Chapman and Hall.
- Celeux, G., E. Diday, G. Govaert, Y. Lechevallier, and H. Ralambondrainy (1988). *Classification automatique des données*. Informatique. Paris: Dunod.
- Dougherty, J., R. Kohavi, and M. Sahami (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 194–202.
- Fisher, W. D. (1958). On grouping for maximum of homogeneity. *Journal of the American Statistical Association* 53, 789–798.
- Fisher, W. D. (1962). Optimal aggregation in multi-equation prediction models. *Econometrica* 30, 744–769.
- Gilula, Z. and A. M. Krieger (1983). The decomposability and monotonicity of Pearson's chi-square for collapsed contingency tables with applications. *Journal of the American Statistical Association* 78, 176–180.
- Govaert, G. (1984). Classification simultanée de tableaux binaires. In E. Diday et al. (Ed.), *Data Analysis and Informatics* 3, Amsterdam, pp. 223–236. North Holland.
- Govaert, G. (1995). Simultaneous clustering of rows and columns. *Control and Cybernetics* 24(4), 438–458.
- Greenacre, M. (1988). Clustering the rows and columns of a contingency table. *Journal of Classification* 5, 39–51.
- Hirotsu, C. (1983). Defining the pattern of association in two-way contingency tables. *Biometrika* 70, 579–589.
- Jobson, J. D. (1992). *Applied Multivariate Data Analysis*, Volume II: Categorical and Multivariate Methods. New York: Springer-Verlag.
- Monti, S. and G. F. Cooper (1999). A latent variable model for multivariate discretization. In *The Seventh International Workshop on Artificial Intelligence and Statistics*.
- Olszak, M. and G. Ritschard (1995). The behaviour of nominal and ordinal partial association measures. *The Statistician* 44(2), 195–212.
- Ritschard, G. and N. Nicoloyannis (2000). Aggregation and association in cross tables. In Zighed, Komorowski, and Zytkow (Eds.), *Principles of Data Mining and Knowledge Discovery*, pp. 593–598. Berlin: Springer-Verlag.
- Ritschard, G., D. A. Zighed, and N. Nicoloyannis (2001). Maximisation de l'association par regroupement de lignes ou colonnes d'un tableau croisé. *Revue Mathématiques Sciences Humaines* 39(154/155), 81–97.
- Ritschard, G. (2001). Performance d'une heuristique d'agrégation optimale bidimensionnelle. *Extraction des connaissances et apprentissage* 1(4), 185–196.
- Zighed, D. A., S. Rabaseda, R. Rakotomalala, and F. Feschet (1999). Discretization methods in supervised learning. *Encyclopedia of Computer Science and Technology* 40(25), 35–50.