ZigZag, a New Clustering Algorithm to Analyze Categorical Variable Cross-Classification Tables

Stéphane Lallich

ERIC Laboratory, University of Lyon 2
e.mail : lallich@univ-lyon2.fr

Abstract. This paper proposes ZigZag, a new clustering algorithm, that works on categorical variable cross-classification tables. Zigzag creates simultaneously two partitions of row and column categories in accordance with the equivalence relation "to have the same conditional mode". These two partitions are associated one to one and onto, creating by that way row-column clusters. Thus, we have an efficient KDD tool which we can apply to any database. Moreover, ZigZag visualizes predictive association for nominal data in the sense of Guttman, Goodman and Kruskal. Accordingly, the prediction rule of a nominal variable Y conditionally to an other X consists in choosing the conditionally most probable category of Y when knowing X and the power of this rule is evaluated by the mean proportional reduction in error denoted by $\lambda_{Y/X}$. It would appear then that the mapping furnished by ZigZag plays for nominal data the same role as the scattered diagram and the curves of conditional means or the straight regression line plays for quantitative data, the first increased with the values of $\lambda_{Y/X}$ and $\lambda_{X/Y}$, the second increased with the correlation ratio or the R^2.

1 Introduction

Extracting knowledge from categorical data cross-classifications. The development of databases offers to researchers and practitioners a high variety of data in various fields like social and economic sciences, business or biomedical sciences. These data are often issued from categorical variables and more specifically from nominal variables. Statistical methods referring to categorical variables have much extended over the last thirty years[1]. For instance, concerning the cross-classification topic, Goodman and Kruskal[4] developed various prediction rules and association coefficients which are the counterpart of the regression for quantitative variables. In this paper, we present ZigZag, an algorithm that creates a partition of the row categories and column categories according to the logic of predictive association as developed by Guttman, Goodman and Kruskal[4]. ZigZag constitutes a simple and efficient tool in order to synthesize and visualize the associations between two nominal variables and so facilitates the extraction of the useful knowledge resulting from the crossing of two categorical attributes in databases.

Notations. We consider a population of subjects which are described by two categorical variables. Let X and Y denotes these two variables, X having p

J.M. Żytkow and J. Rauch (Eds.): PKDD'99, LNAI 1704, pp. 398–405, 1999.

categories and Y having q categories. The responses of the subjects are presented in a rectangular cross-classification (or contingency) table having p rows and q columns. Most of the time, the population is too large and we are in a situation of sampling. Then, the (ν_{ij}) and (π_{ij}) are unknown. We can only observe the (n_{ij}) and the (p_{ij}). A common model of sampling is the multinomial sampling[1] when the observations result from equal probabilities and independent random draws among the whole population.

	Population (theoretical)	Sample (empirical)
Number of sujects	ν	n
Joint absolute freq. of (x_i, y_j)	ν_{ij}	n_{ij}
Joint relative freq. of (x_i, y_j)	$\pi_{ij} = \nu_{ij}/\nu$	$p_{ij} = n_{ij}/n$

Reading categorical variables cross-classifications. To extract the knowledge included in a cross-classification, we generally begin by taking an interest in the conditional mode of each row and each column. If n_{ij} is maximum in the row i, it means that conditionally to X=x_i, most of the time Y=y_j, which leads us to associate y_j to x_i. At the same time, if n_{ij} is maximum in the column j, it means that conditionally to Y=y_j, most of the time X=x_i, which leads us to associate x_i to y_j. The aim of ZigZag is to systematize and automate this process.

2 Main Topics of the Algorithm ZigZag

Let **X** denotes the set of row categories, and **Y** the set of column categories, with Card **X**=p and Card **Y**=q. We construct simultaneously two partitions of row categories (**X**) and column categories (**Y**) on the basis of the maximum association on rows (conditionally to X) and on columns (conditionally to Y). Then we join these two partitions one class to one class in order to obtain row-column clusters.

Best column criterion. For each row category x_i, i=1, 2, ...p, we associate the category $y_{j(i)}$, where j(i)$\in${1,2,...,q}, which represents the mode of the row. Then, the value $n_{i,j(i)}$ is the maximum value of the i^{th} row of (n_{ij}). So we define an application c from **X** to **Y** which is necessarily not onto if p<q, or necessarily not one to one if p>q. The graph of this application c constitutes a bipartite graph denoted by G_c.

Best row criterion. For each column category y_j, j=1, 2, ..., q, we associate the row category $x_{i(j)}$, where i(j)$\in${1, 2,..., p}, which represents the mode of the column. Then, the value $n_{i(j),j}$ is the maximum value of the j^{th} column of (n_{ij}). So we define an application r from **Y** to **X**, which is necessarily not one to one if p<q, or necessarily not onto if p>q. The graph of this application r constitutes a bipartite graph denoted by G_r having the same nodes as G_c.

Strong pattern. If we merge the graphs G_c and G_r, while distinguishing each type of edge (for example with a solid line for c and a dotted line for r), we obtain a bipartite graph G having p+q edges. A pair of nodes (i, j) is relied

by two edges at most. If there are two edges, it is necessary that they be of a different nature, one with a solid line, the other with a dotted line. We will consider such a couple row-column as a strong pattern of the bipartite graph: each member is the image of the other one through the relationship nearest column - nearest row expressed by the graph. The corresponding joint absolute frequency n_{ij} is the maximum for row i and column j.

Whole partition resulting from strong pattern. When the pair consisting in a category of one variable and the associated category of the other variable is not a strong pattern, a chain of "nearest neighbors" appears, like i_1, j_1, i_2, j_2, ..., where j_h is the nearest column of i_h, while i_{h+1} is the nearest row of j_h. Necessarily, each chain ends up by a strong pattern, which consists in reciprocal nearest neighbors. Then, we seek for each row i, i = 1, 2, ..., p, and for each column j, j = 1, 2, ..., q, to which chain it belongs and by which strong pattern its chain ends up. The relation R "to be associated with the same strong pattern" defines an equivalence relation on the nodes of G as well as on the traces G$\frown$**X** et G$\frown$**Y**. The equivalence classes modula R are the connected components of the graph G. Their intersections with **X** and **Y** constitute two partitions of **X** and **Y** joined one class to one class.

ZigZag algorithm. ZigZag has been implemented in Delphi[5] and is now available on the Web. The algorithm is applied in two stages. Firstly, we create the table indicating the nearest column category of each row category and the nearest row category of each column category. Then, using this table, the different chains of nearest neighbors are built and the corresponding graph is represented. The categories which belong to chains ending up by the same reciprocal nearest neighbor's pair constitute a connected component of the graph and define a row-column cluster.The number of clusters is equal to the number of strong patterns, which is comprised between 1 (in case of independence) and min {p, q}, (in case of functionnal dependence).

3 Application to Patents Data

We have created ZigZag on the occasion of a work dealing with French firms patenting in the US over the period 1985-90, using the SPRU-LESA database[2]: each patent granted by the US Patent Office to French based firms and institutions is described by the industrial sector in which it is produced and the technological field of the patent application. To visualize this table, we first used the program AMADO[3] to permute the rows and the columns of data matrix in order to reveal the underlying structure of the matrix (Fig. 1). We also considered using classification of the rows or columns of the table, but we thought that it dealt with the rows and columns in an assymetric manner. In fact, we needed to define real techno-industrial clusters gathering both industrial sectors and technological fields which are strongly related. As a result, we conceived ZigZag, of which the resulting mapping is given below (Fig. 2).

It appears from this mapping that there are ten techno-industrial clusters collecting 41 % of the total number of patents. Among these clusters three are

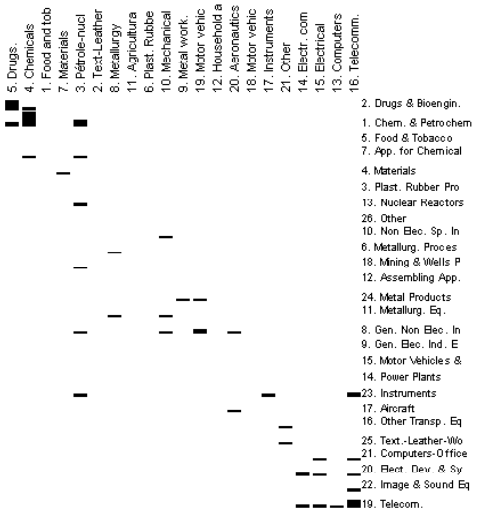

Fig. 1. Bertin's matrix for patent's data.

complex ones, "Chemicals", "Electronical" and "Mechanical Transports Equipment" and represent 33 % of the total patents.

4 Visualization of Predictive Association

ZigZag visualizes predictive association for nominal data in the sense of the λ coefficient proposed by Guttman, Goodman and Kruskal[4].

4.1 Predictive Association for Categorical Data

We recall that it is quite easy to predict Y knowing X when X and Y are quantitative variables. We first use a statistical software to plot the scattered diagram and the straight regression line or the curve of conditional means, increased with the coefficient of correlation or the correlation ratio. Starting from this diagram, it is possible to choose the most appropriate function of X in order to explain Y and to calculate the prediction of Y for a given value of X.

When X and Y are nominal variables, one generally uses the strategy of P.R.E (Proportional Reduction in Error) coefficients of association proposed by Goodman and Kruskal[4]. Firstly, we have to define a prediction rule which can be applied "a priori", without any information about X, and "a posteriori", when knowing X. Then, we calculate the risk of error associated with the "a priori" rule and the mean error risk associated with the "a posteriori" rule.

The P.R.E coefficients measure the proportional mean reduction in error of the prediction rule. The most common coefficients are the λ of Guttman, Goodman and Kruskal, the U of Shannon and the τ of Goodman et Kruskal. We can present them as particular cases of proportional mean reduction in diversity of the Y distribution, as in Lallich[6].

The easiest of these three coefficients is the λ coefficient. The corresponding prediction rule is the optimal one to predict a unique category of the dependent variable: "always guess the modal class observed in the data". The "a priori" rule predicts the modal category of the marginal distribution of the dependent variable, whereas the " a posteriori " rule predicts the modal category of the conditional distribution. If Y is the dependent variable, the error risk of the "a priori" prediction rule is $r_0 = 1 - \max\{\pi_{+j}; j = 1, 2, ...q\}$.The mean error risk of the "a posteriori" prediction rule is:

$$r_1 = 1 - \sum_{i=1}^{p} \pi_{i+} \max\{\pi_{j/i}; j = 1, 2, ...q\}$$

So, the expression of $\lambda_{Y/X}$ is:

$$\lambda_{Y/X} = \frac{r_0 - r_1}{r_0} = \frac{\sum_{i=1}^{p} \pi_{i+} \max\{\pi_{j/i}; j = 1, 2, ...q\} - max\{\pi_{+j}; j = 1, 2, ...q\}}{1 - \max\{\pi_{+j}; j = 1, 2, ...q\}}$$

The extreme value $\lambda_{Y/X}$ =0 only is a necessary condition of independence. On the contrary, $\lambda_{Y/X}$=1 is a necessary and sufficient condition of functional dependence: knowing X you know Y. In the same manner, when considering X as the dependent variable, we calculate $\lambda_{X/Y}$, the other asymmetrical coefficient. Furthermore, one can calculate the symmetrical coefficient λ_{XY}, defined as the mean of the asymmetrical coefficients weighted by their "a priori" error risks.

$$0 \leq Min\{\lambda_{Y/X}; \lambda_{X/Y}\} \leq \lambda_{XY} \leq Max\{\lambda_{Y/X}; \lambda_{X/Y}\} \leq 1$$

If λ_{XY} =0, then $\lambda_{Y/X}$=$\lambda_{X/Y}$=0, but that is only a necessary condition of independence. On the contrary,λ_{XY}=1 implies $\lambda_{Y/X}$=$\lambda_{X/Y}$=1, which is a necessary and sufficient condition of double functional dependence, requiring p=q.

In the case of multinomial sampling, the (p_{ij}) are maximum likelihood estimators of the (π_{ij}) and they are asymptotically normal. Thus the sample association coefficient, denoted by $L_{Y/X}$, is the maximum likelihood estimator of $\lambda_{Y/X}$ and is asymptotically normal. So, it is possible to estimate the asymptotic variance of $L_{Y/X}$ by applying the delta method[4].

In brief, the mapping furnished by ZigZag plays for nominal data the same role as the one played by the scattered diagram with the curves of conditional means or the straight regression line for quantitative data, the first increased with the values of $L_{Y/X}$ and $L_{X/Y}$, the second increased with the correlation ratio or the R^2. In the following, we present one example of prediction for each type of data.

4.2 Usual Approach to Predict Quantitative Variables

To illustrate the usual approach to predict quantitative variables, we use the artificial data that are mentioned below, where X is a discrete variable and Y a continuous one.

X	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3
Y	3.9	3.0	2.8	2.4	2.1	1.4	1.8	2.2	3.7	6.2	5.7	7.3	5.1	5.3	5.0	4.7	6.1	7.5	7.7	9.8

X	3	3	3	3	3	3	4	4	4	4	4	4	4	4	5	5	5	5	5	5
Y	7.5	7.4	5.9	7.3	6.3	8.7	9.1	10.0	8.2	8.1	8.4	10.0	9.2	8.5	12.0	11.0	9.5	10.0	9.8	11.0

Commonly used statistical softwares first draw the scattered diagram and the regression curve which suggest the nature of the relation between X and Y (logarithmic relation in our example, cf. Fig. 3). They then allow to compute the regression equation and the corresponding determination coefficient ($Y = 4.52\ LnX + 2.82$ and $R^2 = 0.86$).

4.3 Counterpart Approach Proposed to Predict Qualitative Variables

To illustrate the counterpart approach that we propose to predict qualitative variables, we use a cross-classification table adaptated from Benzecri. In this classification, row categories are cigarette's brands, while column categories are qualities that anybody can associate with each brand name. ZigZag creates a two ways predictive classification that indicates us which is the most probable top of mind quality knowing the name of the brand, and which is the most probable name of brand associated with each quality (Fig. 4). The different λ coefficients values measure the prediction quality.

Cigarette \ Qualité	1	2	3	4	5	6	7	8	9	10	11
Orly	1	20	9	1	4	3	11	4	9	9	7
Alezan	2	9	23	3	33	9	9	4	12	3	5
Corsaire	14	1	1	15	7	1	1	32	23	9	2
Directoire	38	11	15	15	8	7	17	2	4	8	7
Ducat	18	10	7	6	3	7	4	6	7	4	11
Fontenoy	10	9	11	5	6	5	21	0	13	2	2
Icare	9	1	6	12	6	12	6	9	5	6	6
Zodiaque	5	1	2	18	4	9	1	7	5	8	11
Pavois	9	20	7	4	5	6	5	3	10	1	9
Cocker	4	9	12	25	15	9	4	10	5	6	24
Escale	0	7	3	2	3	6	5	12	13	23	10
Hôtesse	1	12	17	2	3	13	27	7	9	33	5

5 Conclusion

In this paper, a new KDD algorithm has been introduced which allows us to explore significant categorical data cross-classifications through graphical display. ZigZag creates clusters which classify row and column categories according

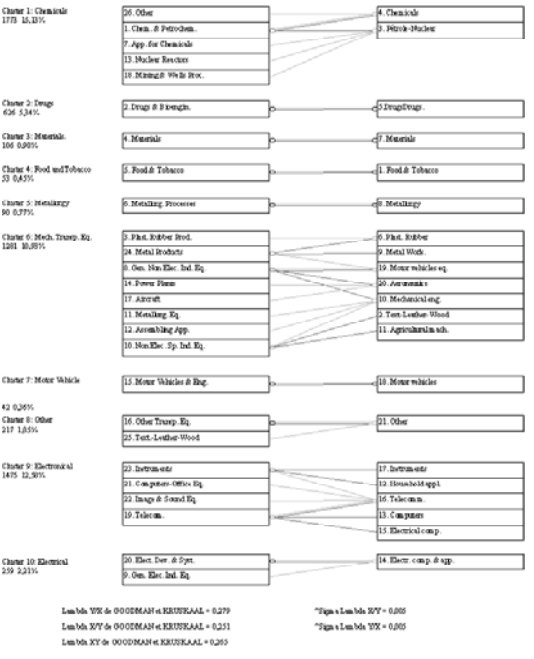

Fig. 2. ZigZag mapping for patent's data

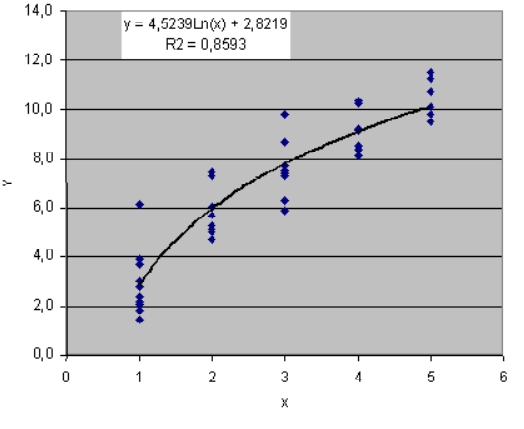

Fig. 3. Scattered diagram for quantitative variable prediction.

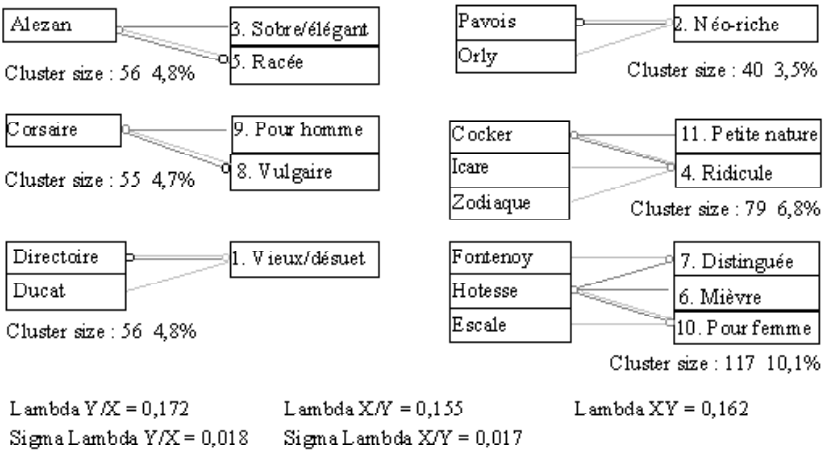

Fig. 4. ZigZag mapping for cigarette's brand data.

to a very simple criterion: "a category is linked to another if it is the most probable conditionally to the latter". Thus, we obtain a visualization of the information contained in the table in terms of predictive association, in the sense of Gutmann, Goodman and Kruskal. The quality of this information is evaluated through $L_{Y/X}$ and $L_{X/Y}$, the two predictive association coefficients created by these authors. As thousands of tables can be computed in many databases, we propose to begin by using $L_{Y/X}$ or $L_{X/Y}$ in order to select the most significant tables. Then, ZigZag is a very efficient tool enabling to synthesize and summarize the knowledge included in these tables.

References

1. Agresti A. (1990), Categorical Data Analysis, John Wiley, New-York.
2. Bergeron S., Lallich S., Le Bas C., (1998), Location of Inovative Activities and Technological Structure in the French Economy, 1985-90: Some Evidences from U.S patenting, Research Policy, 26, pp. 733-751.
3. Chauchat J.H., Risson A. (1998), "Bertin's Graphics and Multidimensional Data Analysis", in Blasius J., Greenacre M. (1998), "Visualization of Categorical Data", Academic Press,
4. Goodman L.A., Kruskal W.H.(1954), Measures of Association for Cross-Classifications I, JASA, 49, pp. 732-764.
5. Guillien F. (1998), Mise en oeuvre de ZigZag sous Delphi: Datamix, Mémoire de Maîtrise Sciences Economiques, Université Lumière Lyon 2.
6. Lallich S. (1999), Concept de diversité et association prédictive, SFDS 99, Grenoble.
7. Rakotomalala R., Lallich S. (1998), Handling Noise with Generalized Entropy of Type Bêta in Induction Graphs Algorithms, JCIS '98, Duke, USA.