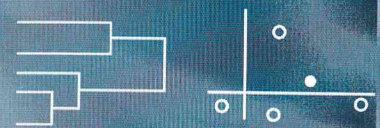
STUDIES IN CLASSIFICATION,
DATA ANALYSIS,
AND KNOWLEDGE ORGANIZATION

Data Analysis and Classification

Francesco Palumbo
Carlo Natale Lauro
Michael J. Greenacre
Editors



Studies in Classification, Data Analysis, and Knowledge Organization

For further volumes:

Francesco Palumbo · Carlo Natale Lauro
Michael J. Greenacre
Editors

Data Analysis and Classification

Proceedings of the 6th Conference of the Classification and Data Analysis Group of the Società Italiana di Statistica

Springer

Editors

Professor Francesco Palumbo
Department of Institution
in Economics and Finance
Università di Macerata
Via Crescimbeni, 20
62100 Macerata
Italy
francesco.palumbo@unimc.it

Professor Carlo Natale Lauro
Department of Mathematics and Statistics
Università Federico II di Napoli
Via Cinthia - Complesso Universitario
di Monte Sant'Angelo
80126 Napoli
Italy
clauro@unina.it

Professor Michael J. Greenacre
Department of Economics and Business
Universitat Pompeu Fabra
Ramon Trias Fargas, 25–27
08005 Barcelona
Spain
michael.greenacre@gmail.com

ISSN 1431-8814
ISBN 978-3-642-03738-2 e-ISBN 978-3-642-03739-9
DOI: 10.1007/978-3-642-03739-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2009936001

Cover design: SPi Publisher Services

Printed on acid-free paper

Preface

This volume contains revised versions of selected papers presented at the biennial meeting of the Classification and Data Analysis Group (CLADAG) of the Italian Statistical Society, which was held in Macerata, September 12–14, 2007. Carlo Lauro chaired the Scientific Programme Committee and Francesco Palumbo chaired the Local Organizing Committee.

The scientific programme scheduled 150 oral presentations and one poster session. Sessions were organised in five plenary sessions, 10 invited paper specialised sessions and 24 solicited paper sessions. Contributed papers and posters were 54 and 12, respectively.

Five eminent scholars, who have given important impact in the Classification and Data Analysis fields, were invited as keynote speakers, they are *H. Bozdogan*, *S.R. Masera*, *G. McLachlan*, *A. Montanari*, *A. Rizzi*.

Invited Paper Specialised Sessions focused on the following topics:

- Knowledge extraction from temporal data models
- Statistical models with errors-in-covariates
- Multivariate analysis for microarray data
- Cluster analysis of complex data
- Educational processes assessment by means of latent variables models
- Classification of complex data
- Multidimensional scaling
- Statistical models for public policies
- Classification models for enterprise risk management
- Model-based clustering

It is worth noting that two of the ten specialised sessions were organised by the French (*Classification of complex data*) and Japanese (*Multidimensional scaling*) classification societies. The SPC is grateful to professors Okada (Japan) and Zighed (France), who took charge of the Japanese and French specialised session organisation, respectively. The SPC is grateful to the Italian statisticians who actively cooperated in the organisation of the specialised and solicited sessions: they were mainly responsible for the success of the conference.

An Algorithm for Earthquakes Clustering Based on Maximum Likelihood

Giada Adelfio, Marcello Chiodi, and Dario Luzio

Abstract In this paper we propose a clustering technique set up to separate and find out the two main components of seismicity: the background seismicity and the triggered one. We suppose that a seismic catalogue is the realization of a non homogeneous space–time Poisson clustered process, with a different parametrization for the intensity function of the Poisson-type component and of the clustered (triggered) component. The method here proposed assigns each earthquake to the cluster of earthquakes, or to the set of independent events, according to the increment to the likelihood function, computed using the conditional intensity function estimated by maximum likelihood methods and iteratively changing the assignment of the events; after a change of partition, MLE of parameters are estimated again and the process is iterated until there is no more improvement in the likelihood.

1 Introduction

A basic description of seismic events provides the distinction of earthquakes in foreshocks, aftershocks, mainshocks and isolated events. A cluster of earthquakes is formed by the main event of each sequence, its foreshocks and its aftershocks, that could occur before and after the mainshock, respectively. Isolated events are spontaneous earthquakes that do not trigger a sequence of aftershocks and because of this characteristic, space–time features of principal earthquakes (main and isolated events) are close to those of a Poisson process that is stationary in time, since the probability of occurrence of future events is constant in time irrespectively of the past activity, even if nonhomogeneous in space. Therefore, the seismogenic features controlling the kind of seismic release of background and clustered seismicity are not similar (Adelfio et al. 2006b), and to describe the seismicity of an area in space, time and magnitude domains, sometimes it is useful to study separately the features

G. Adelfio (✉)
Department of Statistical and Mathematical Sciences, University of Palermo,
viale delle Scienze, ed 13, 90128, Palermo, Italy
e-mail: adelfio@

An Algorithm for Earthquakes Clustering Based on Maximum Likelihood

Giada Adelfio, Marcello Chiodi, and Dario Luzio

Abstract In this paper we propose a clustering technique set up to separate and find out the two main components of seismicity: the background seismicity and the triggered one. We suppose that a seismic catalogue is the realization of a non homogeneous space–time Poisson clustered process, with a different parametrization for the intensity function of the Poisson-type component and of the clustered (triggered) component. The method here proposed assigns each earthquake to the cluster of earthquakes, or to the set of independent events, according to the increment to the likelihood function, computed using the conditional intensity function estimated by maximum likelihood methods and iteratively changing the assignment of the events; after a change of partition, MLE of parameters are estimated again and the process is iterated until there is no more improvement in the likelihood.

1 Introduction

A basic description of seismic events provides the distinction of earthquakes in foreshocks, aftershocks, mainshocks and isolated events. A cluster of earthquakes is formed by the main event of each sequence, its foreshocks and its aftershocks, that could occur before and after the mainshock, respectively. Isolated events are spontaneous earthquakes that do not trigger a sequence of aftershocks and because of this characteristic, space–time features of principal earthquakes (main and isolated events) are close to those of a Poisson process that is stationary in time, since the probability of occurrence of future events is constant in time irrespectively of the past activity, even if nonhomogeneous in space. Therefore, the seismogenic features controlling the kind of seismic release of background and clustered seismicity are not similar (Adelfio et al. 2006b), and to describe the seismicity of an area in space, time and magnitude domains, sometimes it is useful to study separately the features

G. Adelfio (✉)
Department of Statistical and Mathematical Sciences, University of Palermo,
viale delle Scienze, ed 13, 90128, Palermo, Italy

of *independent* events and *triggered* ones. Indeed, to estimate parameters of phenomenological laws useful for the description of seismicity, a reasonable definition of "earthquake cluster" is required; furthermore the prediction of the occurrence of large earthquakes (related to the assessment of seismic risk in space and time) is complicated by the presence of clusters of aftershocks, that are superimposed to the background seismicity, according to some (unknown) mixing parameter, and shade its principal characteristics. For these purposes the preliminary subdivision of a seismic catalog in background seismicity (represented by isolated events, that do not trigger any further event, and the mainshock of each seismic sequence) and clustered events, is sometimes required. At this regard, a seismic sequences detection technique is presented; it is based on MLE of parameters that identify the conditional intensity function of a model describing seismic activity as a clustering-process, which represents a slight modification of the ETAS model (Epidemic Type Aftershocks-Sequences model; Ogata, 1988, Ogata et al., 2004). Diagnostics for these models is discussed in Adelfio and Chiodi (2008).

In Sect. 2 conditional intensity function of point processes is introduced, focusing on the description of ETAS model and related models. In Sect. 3 the features of the proposed method are defined. Finally in Sect. 4 an example of application is proposed and some conclusive remarks for future works are reported.

2 Conditional Intensity Function of the Clustering Procedure

A seismic catalogue, assumed as realization of a space–time point process, contains information about seismic events occurred in a region, in a given time interval. In particular, given a seismic catalogue of n events, the i-th row of the catalogue gives quantitative information about the estimated latitude, longitude and depth (x_i, y_i, z_i), the time of occurrence (t_i) and the magnitude (m_i) of the seismic event U_i, $(i = 1, \ldots, n)$. In this paper the depth z will not be considered, since the high level of its measurement error.

To describe the features of the seismic activity of a space–time area the definition of a conditional intensity function is required. The conditional intensity function of a space–time point process can be defined as

$$\lambda(t, \mathbf{x}|\mathcal{H}_t) = \lim_{dt, d\mathbf{x} \to 0} \frac{E[N([t, t + dt) \times [\mathbf{x}, \mathbf{x} + d\mathbf{x})|\mathcal{H}_t)]}{\ell(dt)\ell(d\mathbf{x})}, \quad (1)$$

where $\ell(x)$ is the Lebesgue measure of x; $\mathcal{H}_t$ is the space–time occurrence history of the process up to time t, i.e. the σ-algebra of events occurring at times up to but not including t; $dt, d\mathbf{x}$ are time and space increments respectively, and $E[N([t, t + dt) \times [\mathbf{x}, \mathbf{x} + d\mathbf{x})|\mathcal{H}_t)]$ is the history-dependent expected value of occurrence in the volume $\{[t, t + dt) \times [\mathbf{x}, \mathbf{x} + d\mathbf{x})\}$. The conditional intensity function completely … process (Daley and Vere-Jones 2003)

the spatial locations (1) supplies a nonhomogeneous Poisson process; a constant conditional intensity provides a homogeneous Poisson process).

For a space–time point process, the log likelihood function is defined by

$$\log L = \sum_{i=1}^{n} \log \lambda(x_i, y_i, t_i) - \int_{T_0}^{T_{max}} \int_{\Omega_{xy}} \lambda(x, y, t) dx\, dy\, dt, \quad (2)$$

where (x_i, y_i, t_i) are the space–time coordinates of the $i-th$ event $(i = 1, 2, \ldots, n)$, $(T_0 - T_{max})$ is the observed period of time and Ω_{xy} is the space region.

2.1 *The ETAS Model*

The conditional intensity function used in our procedure is a variation of ETAS model, a self-exciting point process describing earthquakes catalogs as a realization of a branching or epidemic-type point process. The conditional intensity function of the ETAS model in a point x, y, t, m of the space–time-magnitude domain, conditioned to the space–time occurrence history of the process up to time t, denoted by $\mathcal{H}_t$, is defined by

$$\lambda(x, y, t, m|\mathcal{H}_t) = J(m) \left[\mu(x, y) + \sum_{t_j < t} g(t - t_j) f(x - x_j, y - y_j|m_j) \right], \quad (3)$$

where x_j, y_j, t_j, m_j are the space–time-magnitude coordinates of the observed events up to time t, $J(m)$ is the density of magnitude (Gutenberg and Richter 1944); inside the squared brackets there is the sum of the spontaneous activity $\mu(x, y)$ and the triggered one, given by the product of the time and space (conditioned to magnitude) probability distributions. The main hypothesis of the model states that all events, both a mainshock or an aftershock, have the possibility of generating offsprings.

In ETAS model, background seismicity $\mu(x, y)$ is assumed stationary in time, while time triggering activity is represented by a non stationary Poisson process according to the modified Omori's formula (Utsu, 1961). In this model, the occurrence rate of aftershocks at time t following the earthquake of time τ, is described by

$$g(t - \tau) = \frac{K}{(t - \tau + c)^p}, \quad \text{with } t > \tau \quad (4)$$

with K a normalizing constant, c and p characteristic parameters of the seismic activity of the given region; p is useful for characterizing the pattern of seismicity, indicating the decay rate of aftershocks in time.

In (3), $f(\cdot,\cdot)$ is the spatial distribution, conditioned to magnitude of the generating event; a number of its formulations are proposed in Ogata (1998), where the occurrence rate of aftershocks is related to the mainshock magnitude.

2.2 *Intensity Function for a Particular Clustered Inhomogeneous Poisson Process*

In our procedure, we assume that the seismic catalog is the realization of a clustered inhomogeneous Poisson process, assuming that the events of the background seismicity come from a space–time Poisson process (spatially inhomogeneous) and that among these there is a number k of mainshocks that can generate aftershocks sequences, inhomogeneous both in space and times, as a function of the magnitude of the main event. Differently from ETAS model we do not assume that each event can generate an offspring. Therefore, in our procedure we consider the following intensity function:

$$\lambda(x, y, t; \boldsymbol{\theta}) = \lambda_t \mu(x, y) + K_0 \sum_{\substack{j=1 \\ (t_j < t)}}^{k} g_j(x, y) \frac{\exp[\alpha(m_j - m_0)]}{(t - t_j + c_j)^{p_j}}, \tag{5}$$

where $\boldsymbol{\theta} = (\lambda_t, K_0, c_j, p_j, \alpha)$. In (5) t_j and m_j are time of the first event and magnitude of the mainshock of the cluster j, $g_j(x, y)$ is the space intensity of the cluster j and $\mu(x, y)$ is the background one; K_0 and λ_t are the weights of the clustered seismicity and of the background one, respectively. Background seismicity is assumed stationary in time, while time aftershock activity is represented by the modified Omori formula (Utsu 1961), with parameters c_j and p_j, relating the occurrence rate of aftershocks to the mainshock magnitude m_j, with α measuring the influence on the relative weight of each sequence, m_0 the completeness threshold of magnitude, i.e. the lower bound for which earthquakes with higher values of magnitude are surely recorded in the catalog.

In our approach space intensity, both of background seismicity $\mu(x, y)$ and of each cluster $g_j(x, y)$, $j = 1, \ldots, k$, is estimated by a bivariate kernel estimator: it is computed either using only the independent events (isolated and mainshocks) or points belonging to the cluster, including the mainshock, respectively.

The used bivariate kernel estimator is

$$\hat{f}(x, y) = \frac{1}{n h_x h_y} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h_x}, \frac{y - Y_i}{h_y}\right), \tag{6}$$

where $K(\cdot,\cdot)$ is a generic bivariate kernel function and $\mathbf{h} = (h_x, h_y)$ is the vector [...] the smoothing constant is evalu-

$h_{opt} = 1.06 A n^{-1/5}$, with A = min{standard deviation, range-interquartile/1.34}, which optimizes the estimator asymptotic behavior in terms of mean integrated square error and provides valid results on a wide range of distributions. In the evaluation of (5), different kinds of parametrization are considered to take into account for different assumptions on the seismicity of an area, (e.g. Omori's law parameters of the k clusters can be assumed equal or distinct in each cluster). The choices can be compared at the end of the procedure on the basis of the final likelihood values.

3 The Proposed Clustering Method

In our clustering procedure we assume that a catalog of n events may be partitioned in $k+1$ sets, one relative to the background seismicity and k relative to clusters, according to partition $\mathcal{P}_{k+1}$. Then, on the basis of this partition, three types of events are identified: n_0 isolated points, k mainshocks and n_j points belonging to the j-th cluster ($j = 1, 2, \ldots, k$), where $\sum_{j=0}^{k} n_j = n$. The goal of this method is to find a good partition of events, according to likelihood function maximization, with respect the vector of parameters θ of the intensity function and the partition $\mathcal{P}_{k+1}$, by an iterative procedure. Briefly, given partition $\mathcal{P}_{k+1}$, we compute the estimate $\hat{\theta}$ which maximizes the likelihood function (2) and, on the basis of the estimated value $\hat{\theta}$, we look for a better partition moving single points from their current position to a new subset (a new cluster or the set of main events) such that the likelihood function increases, until a convergence criterion is achieved.

In our approach the likelihood is not complete, but it is conditioned on the assumption that a subset of primary events is known following a version of the original trigger model (see Ogata, 2001). The complete likelihood is simulated alternating the cluster identification and the maximization steps, as in E-M algorithm structure.

3.1 *Finding a Candidate Cluster and Likelihood Changes*

At each iteration s, given $\hat{\theta}^{(s)}$, the cluster r_h which maximizes the conditional intensity function is found for each unit U_h, $(h = 1, \ldots, n)$ either an isolated or a clustered point; approximately this is obtained comparing the k contributions to the sum:

$$\sum_{\substack{j=1 \\ (t_{0_j} < t_h)}}^{k} g_j(x_h, y_h) \frac{\exp\left[\alpha(m_j - m_0)\right]}{(t_h - t_{0_j} + c)^p}$$

and assigning temporally each unit U_h to the cluster r that maximizes

$$\ldots \exp[\alpha(m_r - m_0)] \ldots \tag{7}$$

If the partition changes (from $\mathcal{P}^{(s)}_{k+1}$ to $\mathcal{P}^{*(s)}_{k+1}$) because of a movement of a single unit, we examine the change in the log-likelihood function $\log L(\boldsymbol{\theta};\mathbf{x},\mathbf{y},\mathbf{t},\mathcal{P}^{(s)}_{k+1})$. Schematically, kinds of change of partition are due to three different types of movement of units: unit U_h moves from background seismicity to cluster r (refereed as type A), unit U_h moves from cluster r to the set of background seismicity (type B) and unit U_h moves from cluster r to cluster q (type C).

We compute the variation in the log-likelihood function for each kind of movement (A, B and C) and for each possible change on the current partition induced by the movement of a unit U_h, $h = 1, 2, \ldots, n$, assuming that $\hat{\boldsymbol{\theta}}^{(s)}$ does not change in each iteration.

3.2 The Algorithm of Clustering

The technique of clustering that we propose leads to an intensive computational procedure, implemented by software R (R Development Core Team 2007). The main steps can be summarized as follows:

1. Iteration $s = 1$. The algorithm starts from a partition $\mathcal{P}^{(s)}_{k+1}$ found by a window-based method (similar to a single-linkage procedure) or other clustering hierarchical methods. Briefly, the starting method puts events in a cluster if in it there is at least an event inside a window of δ_s units of space and δ_t units of time. δ_s, δ_t are given as input.
2. Clusters with a minimum fixed number of elements are found out: the number k of clusters is determined.
3. Partition $\mathcal{P}^{(s)}_{k+1}$ is then completed with the set of isolated points, constituted by the n_0 points not belonging to clusters.
4. Estimation of the space seismicity (6) both for background and k clusters.
5. Maximum Likelihood Estimation of parameters: in (5) it is possible to assume either common Omori law parameters c and p over all cluster or varying c_j and p_j in each cluster (this could depend on the available catalog): as default, we consider the second type parametrization. An iterative simplex method is used for the maximization of the likelihood (2). $\hat{\boldsymbol{\theta}}^{(s)}$ is the value of the MLE.
6. Finding a better partition $\mathcal{P}^{(s)}_{k+1}$: for each unit U_h, either an isolated or a clustered point, the *best candidate cluster* r_h is found, according to the rule in (7).
7. Different kinds of movements are tried (type A, B or C, as in Sect. 3.1).
8. Points are assigned to the *best set* of events (best in the sense of likelihood).
9. Points are moved from clusters to background (and viceversa) if their movement increases the current value of the likelihood (2).
10. If no point is moved the algorithm stops, otherwise $\mathcal{P}^{(s)}_{k+1}$ is updated, $s = s + 1$ and the algorithm come back to step 2.

In the last steps (6–9), the likelihood (2), is computed using the current value of $\hat{\boldsymbol{\theta}}^{(s)}$. On the basis of the final partition and the final values of the estimates, the vector of estimated intensities for each point is computed.

4 Application to a Real Catalog and Final Remarks

The proposed method could be the basis to carry out an analysis of the complexity of the seismogenic processes relative to each sequence and to the background seismicity, separately. It has been applied to a catalog of 4,295 seismic events occurred in the Southern Tyrrhenian Sea from February 2000 to December 2005, providing a plausible separation of the different components of seismicity and clusters that have a good interpretability. In Figs. 1 and 2 the found clusters and some of their features are shown. The algorithm identified eight clusters, with a total of 964 events; the remaining 3,331 events were assigned to the background seismicity and the estimated parameters are $\hat{\alpha} = 0.2061$, $\hat{\lambda}_t = 0.000016$ and $\hat{K}_0 = 0.000389$. No relevant dependence of estimated parameters on the magnitude values has been observed.

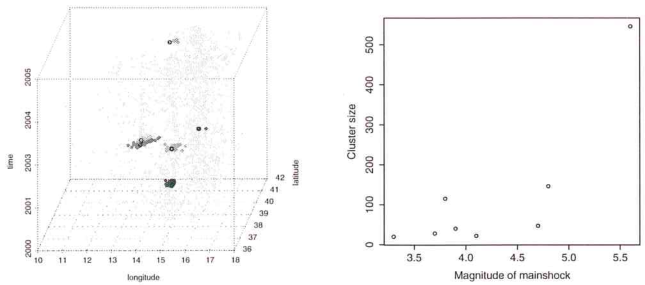
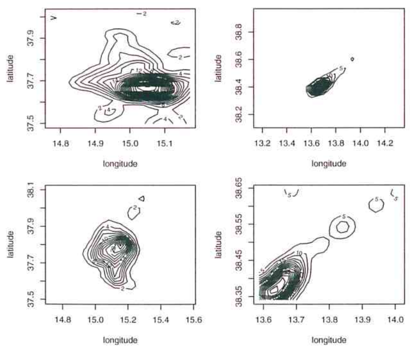


Fig. 1 *On the left:* space–time plot of clusters (*filled circles*) and isolated events (*asterisks*) of the Southern Tyrrhenian catalog from 2000 to 2005. (*Open circle* is used for main event of each cluster.) *On the right:* plot of clusters size vs. mainshocks magnitude



Comparing the current version of the clustering proposed method to its first version (Adelfio et al. 2006a) some extensions have been introduced. In this improved version the moving of points from their current position to a better set (in sense of likelihood) does not require the definition of fixed thresholds and, as described in Sect. 2.2, different kinds of parametrization are introduced allowing to take into account for different assumptions about the seismicity of an area (e.g. Omori law parameters). On the other hand, the optimization steps can be improved in the future, for instance, minimizing the computational burden of the algorithm, reducing the dependence of the convergence of the iterative algorithm on some initial choices.

References

Adelfio, G., Chiodi, M., De Luca, L, & Luzio, D. (2006a). Nonparametric clustering of seismic events. In S. Zani, A. Cerioli, M. Riani, & M. Vichi (Eds.), *Data analysis, classification and the forward search* (pp. 397–404). Berlin: Springer.

Adelfio, G., Chiodi, M., De Luca, L., Luzio, D., & Vitale, M. (2006b). Southern-tyrrhenian seismicity in space–time-magnitude domain. *Annals of Geophysics, 49*(6), 1139–1151.

Adelfio, G., & Chiodi, M. (2008). Second-order diagnostics for space–time point processes with application to seismic events. *Environmetrics*, doi:10.1002/env.961.

Daley, D. J., & Vere-Jones, D. (2003). *An introduction to the theory of point processes* (2nd edition). New York: Springer.

Gutenberg, B., & Richter, C. F. (1944). Frequency of earthquakes in California. *Bulletin of the Seismological Society of America, 34*, 185–188.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association, 83*(401), 9–27.

Ogata, Y. (1998). Space–time point-process models for earthquake occurrences. *Annals of the Institute of Statistical Mathematics, 50*(2), 379–402.

Ogata, Y. (2001). Exploratory analysis of earthquake clusters by likelihood-based trigger models. *Journal of Applied Probability, 38*(A), 202–212.

Ogata, Y., Zhuang, J., & Vere-Jones, D. (2004). Analyzing earthquake clustering features by using stochastic reconstruction. *Journal of Geophysical Research, 109*(B05301), 1–17.

R Development Core Team. (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Utsu, T. (1961). A statistical study on the occurrence of aftershocks. *Geophysical Magazine, 30*, 521–605.

A Two-Step Iterative Procedure for Clustering of Binary Sequences

Francesco Palumbo and A. Iodice D'Enza

Abstract Association Rules (AR) are a well known data mining tool aiming to detect patterns of association in data bases. The major drawback to knowledge extraction through AR mining is the huge number of rules produced when dealing with large amounts of data. Several proposals in the literature tackle this problem with different approaches. In this framework, the general aim of the present proposal is to identify patterns of association in large binary data. We propose an iterative procedure combining clustering and dimensionality reduction techniques: each iteration involves a quantification of the starting binary attributes *and* an agglomerative algorithm on the obtained quantitative variables. The objective is to find a quantification that emphasizes the presence of groups of co-occurring attributes in data.

1 Introduction

Association rules (AR) mining aims to detect patterns of association in large transaction data bases. Transactions are binary sequences recording the presence/absence of a finite set of attributes or items. Let $\mathcal{A} \subseteq \mathcal{I}$ and $\mathcal{B} \subseteq \mathcal{I}$ be two disjoint subsets of the set $\mathcal{I}$ of binary attributes, the expression $(A \Longrightarrow B)$ (to be read *if* A *then* B) represents a general association rule, where $A \in \mathcal{A}$ and $B \in \mathcal{B}$. In the simplest case, both A and B refer to the presence of a single attribute (whereas $\bar{A}$ and $\bar{B}$ refer to the absence).

In other words, an AR is a logical relation: A refers to the antecedent part or *body* and B is termed consequent part or *head*. The association strength of a rule is often measured by the indexes

A. Iodice D'Enza (✉)
Dipartimento di Scienze Economiche e Finanziarie Università di Cassino, Rome