Spatial decision tree-application to traffic risk analysis

Karine Zeitouni, Nadjim Chelghoum

Spatial decision tree-application to traffic risk analysis

Conference Paper · February 2001

DOI: 10.1109/AICCSA.2001.933978 · Source: IEEE Xplore

2 authors:

Karine Zeitouni
Université de Versailles Saint-Quentin
124 PUBLICATIONS **389** CITATIONS
SEE PROFILE

Nadjim Chelghoum
Pierre and Marie Curie University - Paris 6
19 PUBLICATIONS **129** CITATIONS
SEE PROFILE

Spatial Decision Tree
Application to Traffic Risk Analysis

Karine Zeitouni, Nadjim Chelghoum

PRISM Laboratory - University of Versailles
45, avenue des Etats-Unis - F-78035 Versailles Cedex

Tel / Fax : (0)1 39 25 40 46 / (0)1 39 25 40 57

Karine.Zeitouni@prism.uvsq.fr, Nadjim.Chelghoum@prism.uvsq.fr

ABSTRACT. *Spatial data mining fulfills real needs of many geomatic applications. It allows taking advantage of the growing availability of geographically referenced data and their potential richness. This includes the risk analysis linked to a territory such as epidemic risk or traffic accident risk in the road network. This work deals with the method of decision tree for spatial data classification. This method differs from conventional decision trees by taking account spatial relationships in addition to other object attributes. Our approach consists in materializing those spatial relationships (originally implicit) leading to treat them as normal attributes. Then, any conventional method or tool allowing building decision tree could be applied providing naturally a spatial decision tree. Compared to existent approaches, this one is more flexible because no specific algorithm is imposed. Moreover, it considers the organization in several thematic layers that is characteristic of geographical data by distinguishing the intra theme and the inter theme relations (such as the road section contiguity or the proximity between road sections and schools). This method has been tested in the framework of traffic risk analysis.*

KEYWORDS: *Spatial Data Mining, Classification Rules, Decision Tree, Spatial Relationship, Spatial Database.*

1. INTRODUCTION

Nowadays, most government local administrations collect and/or use geographical databases on the road accidents, on the road network and sometimes on the vehicle flow and sometimes on the mobility of inhabitants. In addition, other databases provide additional information on the geographical environment - thematic layers[1] - like administrative boundaries, buildings, census data, etc. These data contain a mine of useful information for the traffic risk analysis. Conscious so, the institutions concerned by the road safety are studying the application of data mining techniques

[1] A thematic layer is a collection of geographical objects that share the same structure. This allows to selectively use the relevant themes for a specific purpose.

for this analysis task. There was a first study (in TOPASE project) aiming identifying and at predicting the accident risk of the roads. It used a decision tree that learns from the inventoried accident data and the description of the corresponding road sections. However, this method is only based on tabular data and does not exploit geographical location. Our project aims at taking account of the spatial feature of the accidents and their interaction with the geographical environment. It involves a new field of data mining technology that is spatial data mining. In our previous work, we have implemented some spatial data mining methods such as generalization and characterization. This paper presents our approach to spatial classification and its application to extend TOPASE.

In the following, we will summarise the works related to spatial data mining and classification emphasising the specificity of this domain. Section 3 describes the traffic risk analysis context and the project background, and then it describes the proposed method as well as its experimentation. A discussion is given in section 4, succeeded by the conclusion and the perspectives.

2. RELATED WORKS

Before presenting our method, this section points out the limitations of traditional analysis methods, emphasises the specificity of spatial data mining, then it reviews both basic and spatial decision tree techniques.

2.1. Limitations of conventional analysis

Right now, data analysis in geography has been essentially based on traditional statistics and multidimensional data analysis and does not take into account spatial property [CHA 95? SAN 89]. This analysis is performed by diverse methods, from the most basic in statistics (average, variance, histograms, etc.), to multivariate analysis, more exploratory and based on the factorial analysis, passing by correlation and regression analysis. All those methods apply to quantitative or qualitative data but not to spatial data. Moreover, they consider the individuals as independents. Consequently, the important feature of spatial auto-correlation is ignored.

Some Geographical Information System (GIS) or statistics tools, however, include geostatistic and spatial statistic functions. This is provided notably in Splus Spatial Stat of MathSoft [MAT 98]. Other tools like Spatial Analyst for ArcView of ESRI [URL 1] allow spatialised analysis, i.e. mapping the statistic analysis results. This is not sufficient for real spatial analysis.

Furthermore, spatial database queries constitute another way to spatial analysis. Those queries can use spatial relationship predicates. For instance, the user can query the database for the counties having more than 10000 inhabitants where the accident rate is more than the average rate. One disadvantage of this kind of analysis is that it is much confirmatory than exploratory. Another is the lack, in database systems, of advanced statistic computations and spatial statistic models [LON 99] - such as Moran and Geary indices -. Nevertheless, this approach could serve in the

filtering phase of the knowledge discovery process, i.e. when preparing the dataset on which the analysis will focus.

2.2. Specificity of spatial data mining

Nowadays, the use of spatial data becomes prevalent for decision support, especially encouraged by geo-data providing on the Internet and by the development of geo-coding tools that transform an address field to a spatial location on a map. For example, in geo-marketing, a store can establish its trade area, i.e. the spatial extent of its customers, and then analyse the profile of those customers on the basis of both their properties and the properties related to the area where they live. However, this involves very large spatial databases, which exceeds human capacity to analyse them. It thus seems appropriate to apply knowledge discovery methods like data mining to spatial data.

Spatial data mining (SDM) arises from the needs of such decisional applications to extract from spatial and regular data useful information or knowledge. This knowledge could be implicit regularities, spatial patterns or relations between spatial data and/or non-spatial data.

The main specificity of SDM is the analysis of spatial relationships. Yet the main specificity of geographic data is that observations located near to one another in space tend to share similar (or correlated) attribute values. In other words, properties concerning a particular place are generally linked and explained in terms of the properties of its neighbourhood [TOB 79]. Consequently, spatial relationships - topological ones [EGE 93], as well as metric (i.e. distance based ones) - have a great importance in the analysis process.

SDM also involves very large databases resulting from the assembling of many geo-data sources. Therefor, optimising techniques are necessary to allow reasonable response time despite this data size.

Nevertheless, SDM methods are largely derived from conventional data mining methods as for the spatial classification. This classification could be used to explain, or to predict located phenomena by analyzing the properties of the geographical environment [ZEI 99], for example to explain the occurrences of accidents by the condition of the road or the urban environment.

2.3. Spatial decision tree

The classification (or class identification) provides a logical description that yields the best partitioning of the entered dataset according to one or few attributes (label class). Classification rules constitute a decision tree where each node contains a criterion on an attribute. The leaves contain objects that belong -in majority- to one label class.

In the spatial database approach, classification is seen as an arrangement of

objects using both their properties (non-spatial values) and their neighbours' properties. Those neighbours could be direct neighbours as well as neighbours of neighbours and so on, up to degree N. Let us take as an example the classification of areas by their economic power. Classification rules are described as follows:

High population ^ neighbour = road ^ neighbour of neighbour = airport => high economic power (95%).

The algorithm proposed by Ester et al. [EST 97] is based on ID3 [QUI 86] and uses the concept of neighborhood graph. The limits of this method is that it does not make distinction between thematic layers, it takes into account only one spatial relationship and moreover, it does not give a real classification of the objects. Indeed, the same object could check several conditions and then would be assigned to several classes.

Koperski et al. [KOP 98] propose another method that considers the spatial predicates (like the adjacency), the spatial functions (such as the distance) as well as the non spatial values of other objects having a spatial relationship with the actual object (like the population living at a given distance from the stores). The originality of this method is that it automatically determines relevant predicates and functions. The relevance of the distance, in other words, the maximum size of the geographical extensions either is determined by an expert, or computed starting from a given maximum distance and decreasing it in the way to maximize the informational gain.

However, this algorithm necessitates to transforms all attribute values into predicates, which is a fastidious task. Another limit is that only one property of neighboring objects is checked (for instance park type in close_to (X, park)). This is why it was not adopted here.

3. THE PROPOSED APPROACH

3.1. The application context

The applications covered by spatial data mining are decisional ones, such as geo-marketing, environmental studies, risk analysis, and so on. In our project[2] spatial data mining is applied to traffic risk analysis [HUG 00]. This application will serve as an example to highlight the specificity and the aim of spatial analysis. We first will briefly describe this project, then we will underline the lack of conventional analysis methods.

---

[2] The first stage of the project was founded by the PSIG (Programme Système d'Information Géographique) of CNRS (the french research center) and IGN (the french mapping agency), in collaboration with INRETS (the French national institute for transport and safety research), while the second stage is a collaboration with the "Ministère de l'équipement" in charge of the road safety and a local administration "Conseil Général des Hauts de Seine".

The traffic risk analysis allows identifying the road safety problem in order to propose safety measures. The risk estimation is based on the information on the previous injury accidents collected by police forces. This analysis has been, right now, of statistic nature but does not use the spatial component or the neighbourhood relationships.

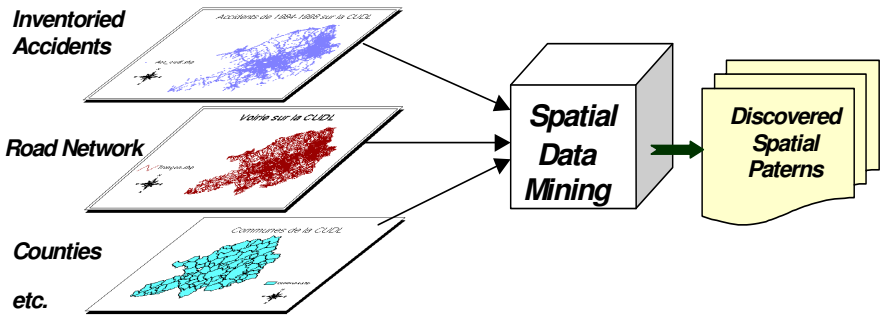

Figure 1: **Spatial Data Mining applied to traffic risk analysis**

Using the accident data, combined to thematic data relating to the road network, the traffic flow, population, buildings, etc., this project aims at deducing relevant risk models to help in traffic safety task. For instance, we should identify regions with a high level of risk and analyse and explain this risk with respect to the geographic neighbourhood. Spatial data mining technology specifically allows for those neighbourhood relationships. This leads to develop a several tools - in spatial data mining domain - by the close collaboration between researchers in risk analysis, in urban geography, in data analysis and in spatial databases.

These project results have been tested using two spatial databases. The one related to Lille City and the urban area around in the north of France, the second concerns the town of Suresnes located in the Hauts de Seines near Paris.

The goal of the present work is to classify the accidents of Suresnes town, by taking account of their neighborhood in three classes of accidents (pedestrians, two-wheeled or only vehicles). The aim is to explain the danger of road sections, not only by their own properties, but also by the geographical environment like the proximity of a school. This leads to extend a decision tree by the spatial dimension.

3.2. Previous work

The former works have focused on descriptive methods rather than predictive methods. This occurs in developing from one hand, spatial statistic methods such as the global and local spatial auto-correlation indices [BLA 98], geographical

clustering [BAN 99], and from the other hand, SDM methods including generalization and characterization [HUG 00]. Besides, and an adaptation of the join index has been proposed to hold spatial relationships and to improve SDM algorithms [ZEI 00]. The method suggested here reuses that work.

3.3. The Method Steps

At the present time, the spatial relationships are initially implicit. They require spatial join operations (i.e. matching data collections according a spatial criterion). The problem is that spatial joins are fastidious and time consuming. Our idea is to make those relationships explicit and to bring back them to semantic properties as well as the other relational references. We have developed spatial join indices [ZEI 00], which pre-compute any spatial join and then improve their performance. The method consists in three stages:

(a) Data cleaning by:

1- eliminating the detail: this step is based on the generalization method previously developed in the project. [ZEI 99]

2- eliminating information (attributes) useless for the analysis: the combination of projection and selection of the relational database model make this.

(b) Retrieving spatial relationships: a spatial join using the spatial join index makes this. Our example uses a distance based join between the thematic layers "accidents" and "constructions".

(c) Building the decision tree: any tool that builds regular decision trees can be used here.

(d) Interpreting the result to extract the rules relevant

3.4. Experimentation

After the stages (a) and (b), the input table of the decision tree is the following:

resultat1.alp - Enregistrements

ID	VID_ACC	NATU_ACCLI	ACCLIB	VID_SEGM	P_FEUX	P_PPAF	P_PPSF	P_TC	DIST_MOYEN	CODE_NATUR	DETAIL_NAT
1	4	autres bles	autres	SEG01250	oui	oui	non	oui	0	autre	Enceinte
2	7	piÙtons ble	piÙtons	SEG00335	non	non	oui	non	92.6575	autre	Enceinte
3	11	piÙtons ble	piÙtons	SEG00272	oui	oui	oui	non	51.2114	admin	Mairie
4	17	2 roues ble	2 roues	SEG00250	oui	oui	oui	non	48.2606	ecole	Enseignement
5	22	2 roues ble	2 roues	SEG00280	non	non	oui	non	3.92882	admin	GendarmerieGendarmerie

Figure 2: Input Data

In this example, the label class corresponds to the Acclib attribute and stands for whom is involved in the actual accident: "piétons" for pedestrians, "2 roues" for two-

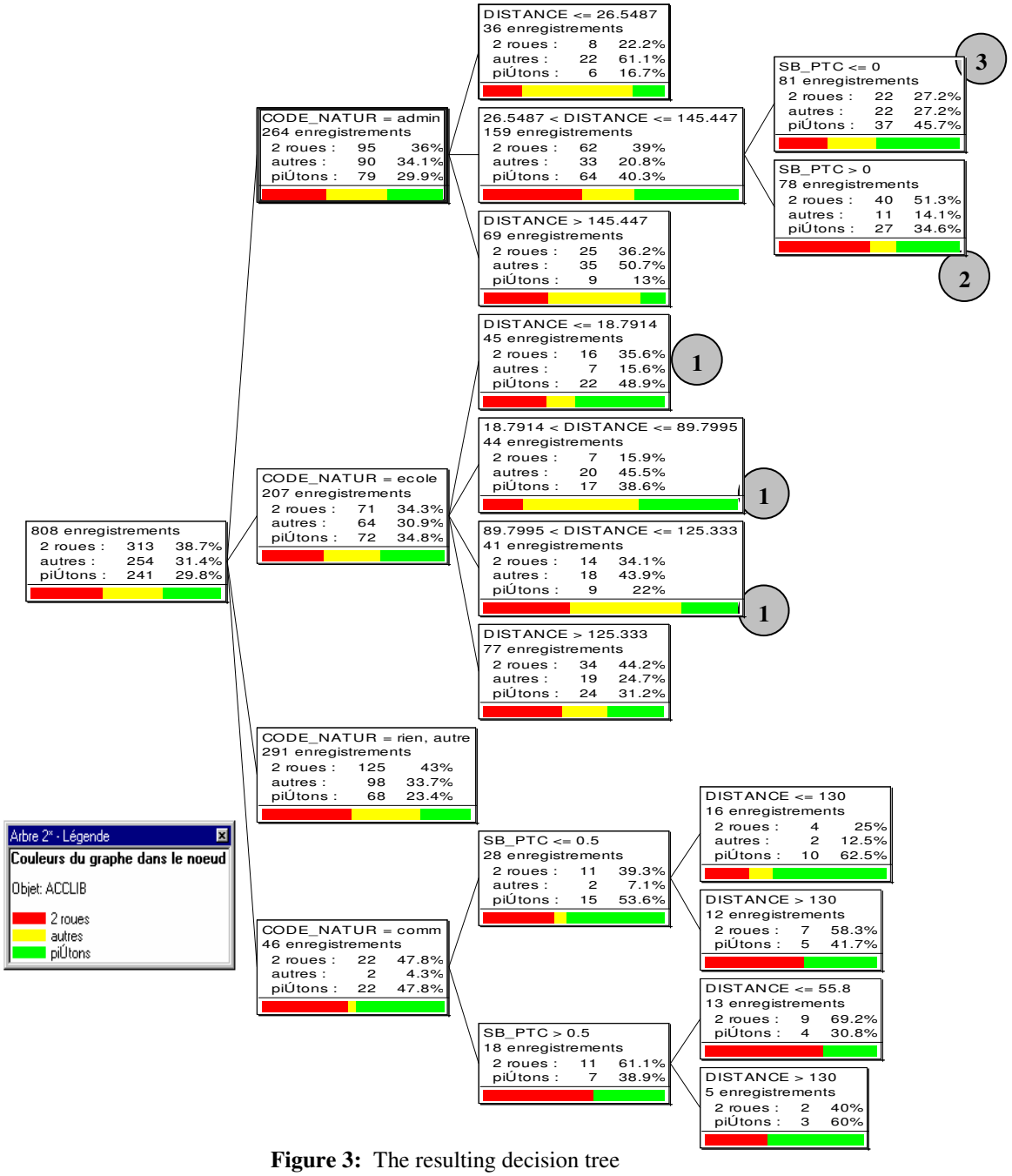

Figure 3: The resulting decision tree

wheeled or "autres" for only vehicles. The predictive attributes describe, from one hand the road section locating the accident like the presence of traffic light,

pedestrian crossing, or bus stop (P_feux, P_ppaf, P_TC) and, from the other hand the nature of the constructions around (code_nature) and their distance to the accident location (dist_moyen).

The decision tree obtained (stage (c)) is given in figure 3. In the interpretation phase (d), we can underline the following rules focusing in the pedestrian class. Node 1 of the figure shows that the amount of accidents involving pedestrians decreases when the section distance to schools increases. There is less pedestrian accidents when there is a bus stop (Node 2) and more, if there is not (Node 3).

4. DISCUSSION

The current application results show a use case of spatial decision trees. The contribution of this approach to spatial classification lies in its simplicity and its efficiency. It makes it possible to classify objects according to spatial information (using the distance). It allows adapting any decision tree algorithm or tool for a spatial modeling problem. Furthermore, this method contrary to the one proposed in [EST 97], considers the structure of geo-data in multiple thematic layers which is characteristic of geographical databases.

Nevertheless, it raised some new questions to which we will focus our future research.

The first problem is related to the limitation of most data mining methods. In effect, they only operate on a single-table and single-row-example format. If the data are stored over multiple tables, or a table contains examples that are described by several rows, it should be pre-processed to fit in the expected single-table format. This conversion may lose potentially valuable information. Furthermore, it involves a pre-processing overhead. In spatial data mining, it is especially important to provide the multi-tables / multi-rows-examples, in reason precisely of the spatial relationships. In order to overcome this limitation, our method transforms multi-tables in a single-table thanks to the spatial join processed in phase (b). But, each example (an accident and its type) is duplicated, as many times as there exists neighboring constructions. This could affect the classification process and the rightness of its resulting rules. Recently, this problem has been addressed by what is called multi-relational data mining [KNO 99]. This will be valuable especially for spatial data mining purpose.

At the moment, the user should enter spatial criteria. A domain-specialist has probably a precise idea of the relevant criteria but this becomes difficult when the number of themes increases. A second question is *"Can the method automatically determines spatial criteria?",* and *"Is it sufficient to consider those criteria as any attribute? ".*

4. CONCLUSION

This work gives a pragmatic approach to multi-layer geo-data mining. The main idea is to prepare input data by joining each layer table using a given spatial criterion, then applying a standard method to build a decision tree. The most advantage is to demonstrate the feasibility and the interest of integrating neighborhood properties when analyzing spatial objects. The discussion above highlights some advantages and limitations of this approach.

Our future work will focus on adapting recent work in multi-relational data mining domain, in particular on the extension of the spatial decision trees in this direction. Another extension will concern automatic filtering of spatial relationships. We will study of its functional behavior and its performances for concrete cases, which has never been done before.

Finally, the quality of this analysis could be improved by enriching the spatial database by other geographical themes, and by a close collaboration with a domain-specialist in traffic risk analysis. Indeed, the quality of a decision tree depends, on the whole, of the quality of the initial data. The resulting models from incomplete, incorrect or non-relevant data inevitably leads to erroneous results. Reliable data thus constitute the key to success of a decision tree. Furthermore, the data preparation phase by a domain-specialist is essential such as the filtering and encoding. The advantage of the technique of decision trees is to allow the end-user (here the domain-specialist) to evaluate the results without any assistance by an analyst or statistician.

Bibliography

[BAN 99] BANOS A., HUGUENIN-RICHARD F., LASSARRE S., 1999, Detection of traffic accidents clusters in a road network, Poster, International Conference on the Analysis and Interpretation of Disease Clusters and Ecological Studies, Royal Statistical Society, London, December 16-17, http://thema.univ-fcomte.fr/BANOS/Banos-Page.htm

[BLA 98] BLACK W., THOMAS I, "Accidents on Belgium's motorways a network autocorrelation analysis", Journal of Transport Geography, Vol. 6, n° 1, 1998, pp. 23-31

[CHA 95] CHARRE J., Statistique et territoire, Editions GIP Reclus, Collection Espaces modes d'emploi, Montpellier, 1995

[EGE 93] EGENHOFER M.J. and SHARMA J., "Topological Relations Between Regions in R2 and Z2", *Advance in Spatial Databases, 5th International Symposium, SSD'93* p. 316-331. Singapore, June 1993, Springer-Verlag.

[EST 97] ESTER M., KRIEGEL H.-P., SANDER J., "Spatial Data Mining: A Database Approach", *Proc. 5th Symp. on Spatial Databases*, Berlin, Germany, 1997.

[HUG 00] HUGUENIN-RICHARD F, LASSARRE S, YEH L, and ZEITOUNI K, "Extraction de connaissances des bases de données spatiales en accidentologie routière", Troisième journée Cassini, La Rochelle, France, Septembre 2000.

[KNO 99] KNOBBE, A,J, SIEBES, A.VAN DER WALLEN, D.M.G. "Multi-relational Decision Tree Induction, In Proceedings of PKDD' 99, Prague, Czech Republic, Septembre 1999.

[KOP 98] KOPERSKI K., HAN J., and STEFANOVIC N., "An Efficient Two-Step Method for Classification of Spatial Data", In *Proc. International Symposium on Spatial Data Handling (SDH'98)*, p. 45-54, Vancouver, Canada, July 1998.

[LON 99] LONGLEY P. A., GOODCHILD M. F., MAGUIRE D. J., RHIND D. W., Geographical Information Systems - Principles and Technical Issues, John Wiley & Sons, Inc., Second Edition, 1999.

[MAT 98] MATHSOFT INC., "S-Plus for ArcView GIS - Users Guide Version 1.0" and "S-Plus Spatial Stat.", Data Analysis Products Division, Seattle, Washington, April 1998.

[QUI 86] QUINLAN J.R., "Induction of Decision Trees." *Machine Learning* (1), 1986

[SAN 89] SANDERS, L.: L'analyse statistique des données en géographie, GIP Reclus, 1989

[TOB 79] TOBLER W. R., *Cellular geography*, In Gale S. Olsson G. (eds) *Phylosophy in Geography,* Dortrecht, Reidel, p.379-86, 1979.

[URL 1] Web site of ESRI Company "URL http://www.esri.com"

[ZEI 99] ZEITOUNI K., YEH L., "Les bases de données spatiales et le data mining spatial", Revue internationale de géomatique, Numéro spécial "Data mining spatial", Vol. 9, N° 4 (99).

[ZEI 00] ZEITOUNI K,YEH L,AUFAURE M, "Join indices as a tool for spatial data mining", International Workshop on Temporal, Spatial and Spatio-Temporal Data Mining, TSDM 2000 Lecture Notes in Artificial Intelligence n° 2007, Roddick J. F. and Hornsby K., Eds., Springer, pp 102-114, September 12-16, 2000, Lyon, France

[ZIG 00] ZIGHED A., RICCO R., *Graphe d'induction - Apprentissage et Data Mining*, Edition Hermès Sciences, 2000.