Mining association rules in asymmetric data for territorial evolution modeling

Asma Gharbi, Cyril de Runz, Herman Akdag, Sami Faiz

Mining Association Rules in Asymmetric Data For Territorial Evolution Modeling

Asma Gharbi
asma.haj-ali-gharbi@univ-paris8.fr
LIASD, University of Paris 8
Saint-Denis, France

Cyril de Runz
cyril.derunz@univ-tours.fr
BdTln, LIFAT, University of Tours
Reims, France

Herman Akdag
Herman@univ-paris8.fr
LIASD, University of Paris 8
Saint-Denis, France

Sami Faiz
sami.faiz@isa2m.rnu.tn
ISAMM
La Mannouba, Tunisia

ABSTRACT

This work starts from the hypothesis that spatial dynamics and the functions (cover or use) of geographical objects could be, partly, explained or anticipated by the history of their functions and co-localizations changes. Hence, an approach relying on association rules mining for the extraction of explicative/predictive models of territorial evolution is proposed. In order to deal with the asymmetry of the used learning data, we proposed to adapt the supports assignment process for the MSApriori and we also proposed a new multiple minimum support based algorithm called BERA. Applied on study cases from the Corine Land Cover database between 1990 and 2012, the proposed mining methods proved their worth in the management of data imbalance and the generated rules highlight realistic urban dynamics.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**; **Association rules**; • **Applied computing** → *Cartography*;

KEYWORDS

Class Association rules, territorial evolution, multiple minimum supports, MSApriori

ACM Reference Format:
Asma Gharbi, Cyril de Runz, Herman Akdag, and Sami Faiz. 2020. Mining Association Rules in Asymmetric Data For Territorial Evolution Modeling. In *The 35th ACM/SIGAPP Symposium on Applied Computing (SAC '20), March 30-April 3, 2020, Brno, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3341105.3374120

1 INTRODUCTION

A territory evolves, is being built and transformed over time. Thus, our societies need tools to explain and even predict, what will be the future evolutions (urban, rural, etc.). Methods, such as data mining, offer tremendous potentials for the terrotorial evolution modeling. Although effective, most of the extracted models focuse, mainly, on the internal characteristics of the studied areas and often neglects the effect of the spatial and temporal relationships embedded in data (i.e., such property is spatially close to such property, and such property appears before such property on the same location). Our approach attempts, to answer these problems by concentrating on the spatio-temporal relations between the geographical entities. Concretely, it proposes to explore the dependencies between the variables describing: the evolution history of spatio-temporal objects, the history of their co-localizations ; and their future functions. These variables correspond respectively to temporal relationships of functions' succession, spatial relationships of neighborhood, and temporal relationships describing the evolution towards a new land use/cover.

Our global approach consists of two phases. In the first one, time series of vector maps are used to track the evolution of objects and then find their life trajectories. In the second phase these latter are preprocessed to generate learning data to which a rules mining algorithm is applied to find evolution rules (a model for territorial evolution). As the learning base generated is in fact asymmetric, we suggest to adapt the mining process by using several frequency thresholds. The goal here is to be able to extract relevant rules involving very frequent items (items corresponding to neighborhood relationships) and others much less frequent (items corresponding to temporal relationships). In this respect, two proposals are made. The first one consists in adapting MSApriori [4], an extension of Apriori using several minimum supports. The second, corresponds to introducing a new algorithm, called BERA, that based on an iterative filtering process finds transactions composed exclusively of frequent items that, written in a particular form, constitute our target rules.

2 BACKGROUND AND RELATED WORK

Rules-based modeling aim at determining, by means of a set of predefined rules, where a certain land use is likely to occur. One of the challenges of the rules-based approach is to define the best combinations of transition rules when there are many variables to handle. There are two common ways to perform this operation. The first one, is known as the trial and error approach. It consists in trying several combinations of parameters, comparing their simulation results and then trying again until satisfaction [1]. The second way is based on statistical methods (e.g. logistic regression),and

SAC '20, March 30-April 3, 2020, Brno, Czech Republic

ACM ISBN 978-1-4503-6866-7/20/03.
https://doi.org/10.1145/3341105.3374120

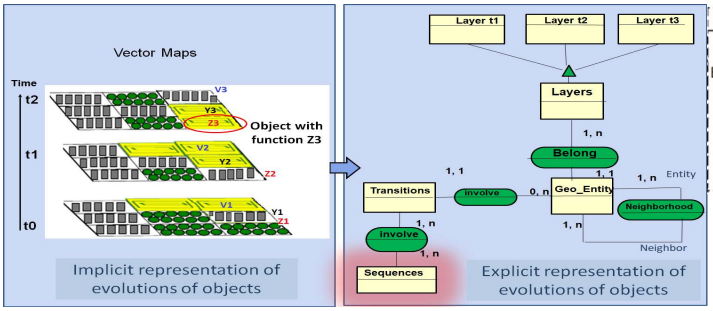

Figure 1: Generating a data model providing life trajectories of georaphic objects.

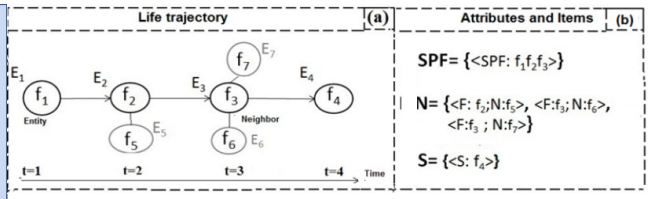

Figure 2: Example of learning data generation.

automated procedures like machine learning algorithms (e.g. artificial neural networks, support vector machine) and data mining approaches. One of the data mining methods is the association rules mining [2]. Consisting in discovering the hidden relationships between different variables in a learning database, our proposed modeling approach (i.e. relying on spatio-temporal relationships for modeling) coincides with association rules mining algorithms. These algorithms, notably Apriori and MSApriori [4], operate in two stages. A first step seeking to regroup the present values (called items) in sets of items, called itemsets. Only those items whose occurrence frequency is greater than a threshold (minSup) or several thresholds (named MIS), in the case of MSApriori, are grouped together to manage possible imbalances in the data. These items are therefore considered frequent, and associated itemsets too. On these frequent itemsets, an approach of generation of association rules is carried out in order to recover the possible rules of co-occurences.

3 EVOLUTION TRACKING

the objective of this phase is to start from a time series of vector maps describing a territory through the states of its objects at some given dates to define a data model explicitly providing their life trajectories. Hence, we consider : **space** as an immutable medium where objects and relationships expressing their spatial configurations are located; **time** as linear in form and we perceive it qualitatively as being a succession of changes of function, form or topology; and a **geographic object** as the product of a spatial dimension (S) (the spatial extent of the object), a temporal dimension (T) (ie, the observation date of the object), and a semantic dimension (F) (the function of the object) [5].

Based on these definitions, we consider that the evolution of an object is characterized by the modification of the value of at least one of its components or dimensions. This change gives rise to the generation of a new version of the object, carrying a new identity and the evolution is described by these different versions. In order to keep the link between them, as proposed in [3],we place the identity tracer on one of the components of the object:the spatial component. Thus, the object is identified by its allocated spatial area and its successors can be identified using spatial overlap and neighborhood queries to reconstruct its life trajectory (see figure 2.a). Neighborhoods are identified according to the topological relationship "touches".

4 DATA REPRESENTATION AND ASSOCIATION RULES MINING

4.1 Learning data representation

Based on the idea that the evolution of a territory can be estimated through the history of the functions of the objects and their co-locations, a life trajectory can be represented by a transaction. It consists of : one item of type (SPF) corresponding to the succession of functions and in our example has the value $< f_1, f_2, f_3 >$; one or more items of type (N) giving the values of the neighbors over time (for instance, the neighborhood relation between the object E_2 of function f_2 and the object E_5 of function f_5, is represented in our example as $< F : f_2; N : f_5 >$; finally, one item of type (S) corresponding to the function of the successor object, in the example $< S : F_4 >$.

These data (transactions and items) are transcribed into the database as a binary table where the items correspond to the columns and the transactions to the rows. A cell is set to 1 if the corresponding item verifies the corresponding transaction (cf. figure 2.a).

4.2 MsApriori-based proposals

Knowing that the generated learning data is unbalanced (very frequent type N items compared to items of types S and SPF), using a single minsup is inappropriate. If it is set too high, only very frequent items (of type N) will appear and if it is set too low, a very large number of items will appear (including rare items of types S and SPf), leading to a combinatorial explosion problem.

MSApriori, which uses several minsup, makes it possible to extract rules containing rarer elements than if we had a single global threshold while reducing the combinatorial complexity if a single low threshold would be defined. However, in MSApriori's assignment process, the expert must define the value of the thresholds based on a subjective assessment of their satisfaction with the models generated. In this article, we seek to minimize the supervision of the threshold assignment process by the use of measures of position.

Quartiles-based method (QuartilesBased): Assuming that they can serve as boundaries between "low" and "high" values, we are interested in the robust measures of position : the median (Q2), the first and the third quartiles(Q1, Q3). Using these measures, the population is divided into four groups of equivalent sizes, corresponding to four semi-open intervals. For each group, its median is assigned as the *MIS* of all the items that belong to it.

Clustering-based method (ClusterBased) : In order to refine the partitioning, we can use clustering algorithms to consider the frequencies of the items and to compare them mutually, in order to group them by similarity of their frequencies. Not wishing to

assume the number of frequency groups, we opted, in the Cluster-Based variant, for the expactation-Maximization (EM) algorithm which does not require this setting.

Integration of the semantics of predicates

Assuming that the semantics (S, SPF, N) of the items may play a role in the process, we proposed a variant of QuartilesBased and a variant of ClusterBased called, respectively, QuartilesBasedSem and ClusterBasedSem.

QuariltesBasedSem : In this variant, we partition the items according to their semantics into 3 groups (SPF item group, item group N, item group S). For each group, the value of the appropriate quartile ($Q1$, $Q2$, $Q3$) is calculated and then assigned as the *MIS* of all its items. For the less frequent attributes (S and SPF), the median ($Q2$) is adopted as *MIS*. For the most frequent attributes (N), the third quartile ($Q3$) is adopted. newline emph ClusterBasedSem : In this variant, the clustering is done on data having for clustering attributes the frequency of the items and their semantics. The median of each group serves as its corresponding MIS.

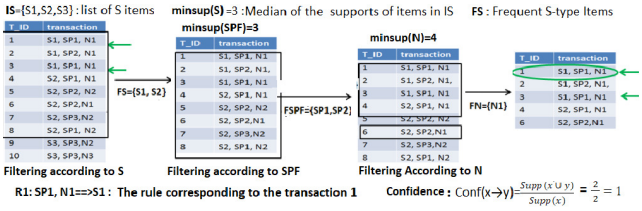

Figure 3: Example illustrating the functioning of BERA.

4.3 BERA

The idea behind BERA algorithm is that rules can be extracted from a set of transactions composed exclusively of frequent items. These transactions are found after an iterative filtering process of the initial transactional database. Each filtering iteration corresponds to a learning attribute (S, SPF, or N). In fact, we start from the idea that each learning attribute represents a different frequency level and, therefore, corresponds to a frequency threshold of its own. Thus, in each filtering iteration: we consider an attribute, we calculate a minsup for all that attribute's items, we identify the frequent items (that correspond to it), and finally we filter the database by deleting any transaction that does not contain one of these already found frequent items. In its operation, BERA generates the rules by starting from their conclusions to determine their premises (cf. figure 3). The transactions in the final dataset are exclusively composed frequent items. Each transaction contains a single S-type item, a single SPF-type item, and several N-type items. The rules are generated as follows: for instance, the , the rule that correponds to the transaction $< s_1, spf_1, n_1 >$ takes this form $spf_1, n_1 \longrightarrow s_1$.

5 RESULTS

Our proposals have been applied to real data from Corine Land Cover for Paris. Four vector maps dating from 1990, 2000, 2006 and 2012 are provided. The first three are used for the generation of the models and the fourth for the evaluation of these. The learning database generated from these vector maps, has 3913 instances,

Table 1: Performances in terms of generation of S and SPF items

	$\frac{\|\{itemsSetSPF\}\|}{\|\{items\}\|}$	elevant rules produced	Stability rules	Transition rules
US : Apriori (minsup=40%)	0	0	0	0
US : MSApriori (β=0.5)	0.086	0	0	0
MS:QuartilesBased	0.550	24816	2469	127
MS:QuartilesBasedSem	0.753	-	-	-
MS:ClusterBased	0.462	6264	6137	127
MS:ClusterBasedSem	0.5	9363	9236	127

190 items (97 N and 93 S and SPF). In table1, the proportion of SPF, N items on the total number of the extracted items shows that all the proposed methods were able to generate S and SPF items. We note that the quartile-based methods outperform those based on clustering, and that taking into account the semantics of the elements in the assignment of minsups, improves these ratios (e.g. 0.55 for QuartilesBased versus 0.75 for QuartilesBasedSem). However, we were unable to extract rules in a reasonable time for QuartilesBasedSem. We distinguish two types of rules : stabiity rules (no change of function between instants t and t-1); and the transition rules where the function of an object changes. We may see in table1 that the three methods for which we could extract rules provide the same number of transition rules but not the same number of stability rules.

6 CONCLUSIONS

This article addresses issues related to the representation of spatio-temporal relationships embedded in raw vector maps, producing rules in a form which is appropriate to our prediction problem, and taking into account rare items that may be useful, by suggesting to adequately specify several frequency thresholds.

Our short-term perspectives consist, mainly, in applying data mining techniques on the generated rules, in order to obtain semantic patterns indicating the most instructive ones.

REFERENCES

[1] Keith C Clarke. 2018. Cellular automata and agent-based models. *Handbook of regional science* (2018), 1–16.

[2] Philippe Fournier-Viger, Jerry Chun-Wei Lin, Bay Vo, Tin Truong Chi, Ji Zhang, and Hoai Bac Le. 2017. A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 7, 4 (2017), e1207.

[3] Tomi Kauppinen and Eero Hyvönen. 2007. *Modeling and Reasoning About Changes in Ontology Time Series.* Springer US, Boston, MA, 319–338. https://doi.org/10.1007/978-0-387-37022-4_11

[4] Bing Liu, Wynne Hsu, and Yiming Ma. 1999. Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 337–341.

[5] Julien Perret, Cyril De Runz, Xavier Rodier, Anne Varet-Vitu, Bertrand Dumenieu, Laure Saligny, Pascal Cristofoli, Bastien Lefebvre, and Eric Desjardin. 2015. Études des dynamiques de l'occupation du sol : questionnement, simplification et limites. *Revue Internationale de Géomatique* 25, 3 (2015), 301–330.