Expert Systems with Applications 42 (2015) 7698–7709

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Flexible propositionalization of continuous attributes in relational data mining

Chowdhury Farhan Ahmed *, Nicolas Lachiche, Clément Charnay, Soufiane El Jelali, Agnès Braud

ICube Laboratory, University of Strasbourg, France

ARTICLE INFO

Article history:
Available online 29 May 2015

Keywords:
Relational data mining
Propositionalization
Numeric attributes
Aggregation
Knowledge discovery

ABSTRACT

In a relational database, data are stored in primary and secondary tables. Propositionalization can transform a relational database into a single attribute-value table, and hence becomes a useful technique for mining relational databases. However, most of the existing propositionalization approaches deal with categorical attributes, and cannot handle a threshold on an attribute and a threshold on the number of objects satisfying the condition on the attribute at the same time. In this paper, we propose a new propositionalization technique called Cardinalization to solve these problems. In order to handle relative numbers, we propose a second variant of our approach called Quantiles which can discretize the cardinality of Cardinalization and achieve a fixed number of features. Therefore, the Quantiles method can be tuned to different deployment contexts. Additionally, we often observe that the best combination of propositionalization and classification methods depends on the new context (e.g., online/incremental learning). One effective solution could be to predict the optimal combination at training time and use it in different deployment contexts. Here we also propose an effective wrapping algorithm, called WPC (Wrapper to combine Propositionalizer and Classifier) to select the best combination of propositionalization and classification methods to address this task. Extensive performance analyses in synthetic and real-life datasets show that our approach is very effective and efficient in relational data mining.

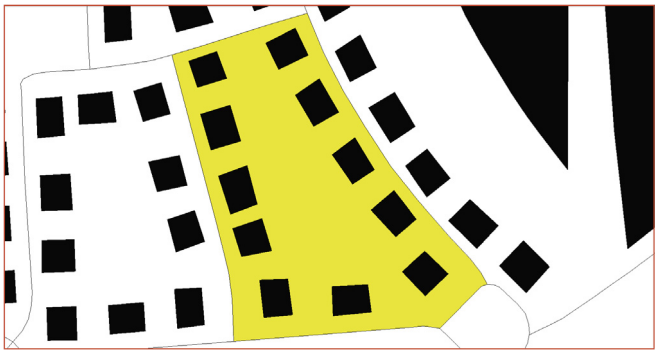
1. Introduction

Data mining techniques discover potentially interesting and useful knowledge from databases which cannot be extracted trivially (Ahmed, Tanbeer, Jeong, & Choi, 2012; Samiullah, Ahmed, Fariha, Islam, & Lachiche, 2014; Zaki & Meira, 2014). Relational data mining (Džeroski & Lavrač, 2001; Kavurucu, Senkul, & Toroslu, 2009; Maervoet, Vens, Berghe, Blockeel, & Causmaecker, 2012) is an important field of data mining which discovers knowledge from relational databases. In a relational database, data are stored in multiple relations/tables and connected through some common keys/fields. The one-to-many relationship is a special kind of link for which each tuple of a primary table may be linked to several tuples of a secondary table. This type of relationship is useful for representing several real-life scenarios, for example customers and purchases in market basket databases, urban blocks and buildings in geographical databases, molecules and atoms in chemical databases, departments and students in university databases, phone numbers and call records in telecommunication databases, and so on. A way of mining these data consists in transforming them into a single attribute-value table. This transformation is called propositionalization (Lachiche, 2010). This paper focuses on propositionalization of relational data involving continuous attributes.

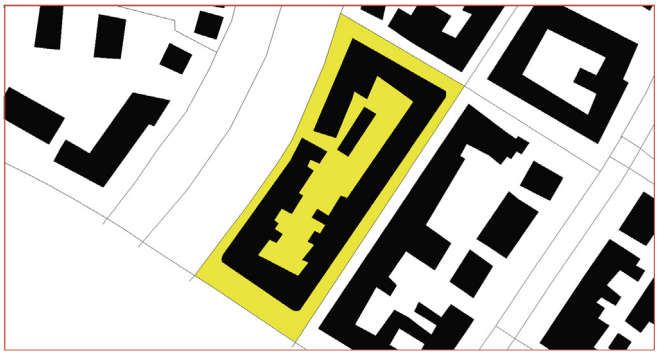
A geographical problem motivated this work. This problem consists in predicting the class of urban blocks (see Fig. 1). The experts have defined 7 classes (Lesbegueries et al., 2009): Continuous urban fabric (city center), Discontinuous urban fabric with individual houses, Discontinuous urban fabric with housing blocks (blocks of flats), Mixed urban fabric (including individual housing and housing blocks), Mixed areas, High density of specialized areas (including industrial, commercial, hospital or scholar buildings), and Low density of specialized areas (containing few or no building). An urban block is characterized only by the geometrical properties of its polygon: area, elongation and convexity. The buildings contained in the urban block are represented as polygons characterized by the same geometrical properties. Density is an additional property of the urban block.

* Corresponding author. Tel.: +33 368 854 577.
E-mail addresses: cfahmed@unistra.fr, farhan@cse.univdhaka.edu (C.F. Ahmed), nicolas.lachiche@unistra.fr (N. Lachiche), charnay@unistra.fr (C. Charnay), soufiane_jelali@yahoo.fr (S. El Jelali), agnes.braud@unistra.fr (A. Braud).

http://dx.doi.org/10.1016/j.eswa.2015.05.053
0957-4174/



(a) Individual houses



(b) Housing blocks

Fig. 1. Geographical example: prediction of the class of an urban block.

Example 1. Tables 1 and 2 represent a sample of a relational database, that we will use as a running example in this article. Here, Tables 1 and 2 present the two urban blocks of Fig. 1 and their buildings, respectively. They also show how data are organized by primary and secondary tables within a relational database.

Discussions with the experts showed that the class depends on conditions about the geometry of buildings and the number, or proportion, of buildings satisfying those conditions. For example, the class "individual houses" mainly depends on the presence of small buildings. Therefore, the learning task consists in determining relevant attributes and their thresholds, as well as the number of those buildings. Existing approaches are not optimized to search at the same time a threshold on an attribute and a threshold on the number of objects satisfying the condition on the attribute. Moreover, they are not suitable for performing propositionalization in different contexts. Motivated by the real-world scenario, in this paper, we propose a new approach for propositionalization to solve these problems.

Furthermore, let us consider a scenario where we are performing online/incremental learning on relational data. At first, we have several training data, enough time and memory to observe the performance of various combinations of propositionalization and classification methods, and choose the best one (e.g., Relaggs and Decision Tree). However, when we are moving towards the next time slots, we might not have the luxury of time to observe the performance of many combinations of propositionalization and classification methods and choose the best one for every time periods. Besides, we might not have enough memory to keep all the combination models to be used for testing onwards. A nice idea would be to take a combination at first and use it over the next time periods without re-checking the performance of other combinations. Can we initially choose the optimal combination of propositionalization and classification methods from the training data whose performance will not mismatch significantly compared to the performance of the actual optimal combinations of upcoming time slots? This gives us the motivation to design an effective wrapping algorithm to choose the optimal combination from training data.

Table 1
The primary table: block.

Idblock	Density	Area	Elong.	Convex.	Class
9601	0.194	6812	0.772	0.921	indiv
9602	0.455	11,119	0.470	0.916	hous
⋮	⋮	⋮	⋮	⋮	⋮

Table 2
The secondary table: building.

Idbuild	Area	Elong.	Convex.	Idblock
4519	122	0.765	1.00	9601
4521	122	0.752	1.00	9601
4528	119	0.948	1.00	9601
4537	112	0.918	1.00	9601
4545	121	0.829	1.00	9601
4556	136	0.739	0.999	9601
4564	115	0.755	1.00	9601
4568	134	0.829	0.999	9601
4579	125	0.745	1.00	9601
4583	98	0.935	0.999	9601
4589	113	0.909	1.000	9601
4231	1669	0.955	0.680	9602
4866	2239	0.772	0.595	9602
4867	229	0.818	0.999	9602
4868	164	0.795	0.936	9602
4869	559	0.451	0.894	9602
4870	205	0.271	0.999	9602
⋮	⋮	⋮	⋮	⋮

The key contributions of our paper are as follows

- We propose a new approach for propositionalization in relational data mining. The first variant of our approach, called Cardinalization, handles a threshold on an attribute and a threshold on the number of objects satisfying the condition on the attribute at the same time. But, it cannot efficiently tackle relative numbers. The second variant of our approach, called Quantiles, can discretize the cardinality of Cardinalization in order to achieve a fixed number of features. Hence, it is quite effective to be applied on relative numbers and suitable for tackling scenarios in different contexts.
- Our approach is very efficient to deal with numeric attributes.
- An effective wrapping algorithm, called WPC (Wrapper to combine Propositionalizer and Classifier), is proposed to choose the optimal combination of propositionalization and classification methods from training data. This optimal combination can be used in several deployment contexts while ensuring a similar performance compared to the actual best combination of that particular context.
- Examples of real-life applications are given to demonstrate the realistic usefulness of our approach.
- Extensive performance study was performed on different synthetic and real-life datasets to show the effectiveness and efficiency of the proposed techniques.

Expert and intelligent systems have good decision-making capabilities in different scenarios. In particular, they can make effective decisions from the built knowledge-base if a model is