Chapter 8
Discovering Interesting Patterns in Numerical Data with Background Knowledge

Szymon Jaroszewicz
National Institute of Telecommunications, Poland

ABSTRACT
The paper presents an approach to mining patterns in numerical data without the need for discretization. The proposed method allows for discovery of arbitrary nonlinear relationships. The approach is based on finding a function of a set of attributes whose values are close to zero in the data. Intuitively such functions correspond to equations describing relationships between the attributes, but they are also able to capture more general classes of patterns. The approach is set in an association rule framework with analogues of itemsets and rules defined for numerical attributes. Furthermore, the user may include background knowledge in the form of a probabilistic model. Patterns which are already correctly predicted by the model will not be considered interesting. Interesting patterns can then be used by the user to update the probabilistic model.

INTRODUCTION
Association rule mining (Agrawal, Imielinski & Swami, 1993) is one of the most important data mining tasks. Initially only simple conjunctions of items were allowed as patterns, but generalizations of the framework to other pattern types such as sequences, trees, graphs etc. have been developed, significantly expanding its applicability. Curiously there has been relatively little effort devoted to generalizing association rules to numerical attributes, despite the practical ubiquity of numerical data.

The main approach to mining numerical data has been discretization (Srikant & Agrawal, 1996). In this approach numerical attributes are split into a number of discrete intervals, after which the data can be mined using standard techniques. Discretization however has several problems. First, discretizing the attributes leads to information loss. Second, each interval contains only a small portion of the data which can lead to statistical estimation problems. The third problem is that relationships between at-
attributes are split among several intervals, so the underlying patterns are harder to spot. Increasing the number of intervals alleviates the first problem, but makes the two remaining ones more acute, so it is not a completely satisfactory solution.

In recent years important progress has been made in the area of mining association rules in numerical data without discretization (Steinbah, Tan, Xiong, Kumar, 2004; Rückert, Richter, Kramer, 2004; Rückert, Kramer, 2006; Achtert, Böhm, Kriegel, Kröger, Zimek, 2006; Besson, Robardet, De Raedt, Boulicaut, 2006; Jaroszewicz 2006; Claders, Goethals, Jaroszewicz 2006; Jaroszewicz, Korzeń 2007; Jaroszewicz, 2008). Those works will be discussed in more detail in the following section.

This chapter presents a method for mining numerical data in the style of association rule mining without discretization. It is possible to discover arbitrary nonlinear relationships through the use of polynomial approximations. Furthermore it extends the method such that the user can provide background knowledge in the form of a probabilistic model. Patterns which are correctly modeled by background knowledge are considered uninteresting and will not be ranked highly. This is a very important improvement as association rule mining algorithms typically produce thousands of patterns which then have to be searched manually. Including background knowledge in the process gives a solid framework for handling interestingness related issues. After interesting patterns have been presented to the user, he/she can update background knowledge based on those patterns, after which a new set of patterns will become interesting. This interactive approach has several advantages such as the understandability of constructed probabilistic models.

BACKGROUND

We will now discuss previous work related to mining association rules in numerical data without discretization. In Rückert, Richter, Kramer, (2004), Georgii, Richter, Rückert, Kramer (2005), Rückert, Kramer (2006) an approach is presented based on finding rules of the type “if a linear combination of some set of attributes exceeds some threshold $a$, than another linear combination of another set of attributes is likely to exceed some threshold $b$”. As sharp thresholds are used, the approach cannot represent functional relationships between attributes, contrary to the approach presented in this Chapter. In Achtert, Böhm, Kriegel, Kröger, Zimek (2006) a method for summarizing clusters of numerical data using linear equations is described. The authors use a clustering algorithm to do the actual pattern discovery, and their approach does not follow the association rule framework.

The idea of Steinbah, Tan, Xiong & Kumar (2004) is closer to our approach. They present a definition of support for numerical data, which does not require discretization. Unfortunately the presented definition of support is not very intuitive, although some interpretation in terms of a lower bound on scalar products is proposed. In similar spirit Calders, Goethals, Jaroszewicz (2006) and Jaroszewicz (2006) presented definitions of support for numerical data based on ranks and polynomials which are easier to interpret.

This work is based on the work Jaroszewicz (2008) where so called minimum variance associations were proposed for mining arbitrary nonlinear relationships in data without discretization. This Chapter extends the method by including background knowledge in the process of selecting interesting patterns. The method will be described in detail in the following section.

There is some similarity between this approach and equation discovery (Dzeroski, Todorovski, 1995), however our approach is much more efficient (see (Jaroszewicz, 2008) for a comparison with the Lagrange equation discovery system) as only a single eigenvalue computation per pattern is required instead of a combinatorial search.

One problem faced by most approaches to
Related Content

The Business Impact of Predictive Analytics
Tilmann Bruckhaus (2007). Knowledge Discovery and Data Mining: Challenges and Realities (pp. 114-138).
www.igi-global.com/chapter/business-impact-predictive-analytics/24904?camid=4v1a

Multiagent Knowledge-Based System Accessing Distributed Resources on Knowledge Grid
www.igi-global.com/chapter/multiagent-knowledge-based-system-accessing/46899?camid=4v1a

Mining Rare Association Rules by Discovering Quasi-Functional Dependencies: An Incremental Approach
Giulia Bruno, Paolo Garza and Elisa Quintarelli (2010). Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection (pp. 131-149).
www.igi-global.com/chapter/mining-rare-association-rules-discovering/36904?camid=4v1a

How to Capitalize Knowledge within Online Communities: An Approach Based on the SECI Model and an Empirical Method of Questioning
www.igi-global.com/chapter/how-to-capitalize-knowledge-within-online-communities/117849?camid=4v1a