Chapter XII
Discovering Knowledge from Local Patterns in SAGE Data

Bruno Crémilleux
Université de Caen, France

Arnaud Soulet
Université François Rabelais de Tours, France

Jiři Kléma
Czech Technical University in Prague, Czech Republic

Céline Hébert
Université de Caen, France

Olivier Gandrillon
Université de Lyon, France

ABSTRACT

The discovery of biologically interpretable knowledge from gene expression data is a crucial issue. Current gene data analysis is often based on global approaches such as clustering. An alternative way is to utilize local pattern mining techniques for global modeling and knowledge discovery. Nevertheless, moving from local patterns to models and knowledge is still a challenge due to the overwhelming number of local patterns and their summarization remains an open issue. This chapter is an attempt to fulfill this need: thanks to recent progress in constraint-based paradigm, it proposes three data mining methods to deal with the use of local patterns by highlighting the most promising ones or summarizing them. Ideas at the core of these processes are removing redundancy, integrating background knowledge, and recursive mining. This approach is effective and useful in large and real-world data: from the case study of the SAGE gene expression data, we demonstrate that it allows generating new biological hypotheses with clinical applications.

INTRODUCTION

In many domains, such as gene expression data, the critical need is not to generate data, but to derive knowledge from huge and heterogeneous datasets produced at high throughput. It means that there is a great need for automated tools helping their analysis. There are various methods, including global techniques such as hierarchical clustering, K-means, or co-clustering (Madeira & Oliveira, 2004) and approaches based on local patterns (Blachon et al., 2007). In the context of genomic data, a local pattern is typically a set of genes displaying specific expression properties in a set of biological situations. A great interest of local patterns is to capture subtle relationships in the data which are not detected by global methods and leading to the discovery of precious nuggets of knowledge (Morik et al., 2005). But, the toughness of extraction of various local patterns is a substantial limitation of their use (Ng et al.,1998; Bayardo, 2005). As the search space of the local patterns exponentially grows according to the number of attributes (Mannila &Toivonen, 1997), this task is even more difficult in *large* datasets (i.e., datasets where objects having a large number of columns).This is typically the case in gene expression data: few biological situations (i.e., objects) are described by ten of thousands of gene expressions values (i.e., attributes) (Becquet et al. 2002). In such situations, naive methods or usual level-wise techniques are unfeasible (Pan et al.,2003; Rioult et al., 2003). Nevertheless, especially in the context of transactional data, the recent progress in constraint-based pattern mining (see for instance (Bonchi & Lucchese, 2006; De Raedt et al., 2002) enable to extract various kind of patterns even in large datasets (Soulet et al., 2007). But, this approach has still a limitation: it tends to produce an overwhelming number of local patterns. Pattern flooding follows data flooding: the output is often too large for an individual and global analysis performed by the end-user. This is especially true in noisy data,such as genomic data where the most significant patterns are lost among too many trivial, noisy and redundant information. Naive techniques such as tuning parameters of methods (e.g., increasing the frequency threshold) limit the output but only lead to produce trivial and useless information.

This paper tackles this challenge. Relying on recent progress in constraint-based paradigm, it presents three data mining methods to deal with the use of local patterns by highlighting the most promising ones or summarizing them. The practical usefulness of these methods are supported by the case study of the SAGE gene expression data (introduced in the next section). First, we provide a method to mine the set of the simplest characterization rules while having a controlled number of exceptions. Thanks to their property of minimal premise, this method limits the redundancy between rules. Second, we describe how to integrate in the mining process background knowledge available in literature databases and biological ontologies to focus on the most promising patterns only. Third, we propose a recursive pattern mining approach to summarize the contrasts of a dataset: only few patterns conveying a trade-off between significance and representativity are produced. All of these methods can be applied even on large data sets. The first method comes within the general framework of removing redundancy and providing lossless representations whereas the two others propose summarizations (all the information cannot be regenerated but the most meaningful features are produced). We think that these two general approaches are complementary. Finally, we sum up the main lessons coming from mining and using local patterns on SAGE data, both from the data mining and the biological points of view. It demonstrates the practical usefulness of these approaches enabling to infer new relevant biological hypotheses.

This paper abstracts our practice of local patterns discovery from SAGE data. We avoid technical details (references are given for in-depth information), but we emphasize the main principles and results and we provide a cross-fertilization of our “in silico” approaches for discovering knowledge in gene expression data from local patterns.