Information Systems 40 (2014) 1–10

Contents lists available at ScienceDirect

Information Systems

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

A new method for mining disjunctive emerging patterns in high-dimensional datasets using hypergraphs

Renato Vimieiro [a,b], Pablo Moscato [a,b,*]

[a] Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia
[b] Hunter Medical Research Institute, Lot 1, Kookaburra Circuit, New Lambton Heights, NSW 2305, Australia

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form
17 September 2013
Accepted 18 September 2013
Recommended by: D. Shasha
Available online 2 October 2013

Keywords:
Emerging patterns
Contrast pattern mining
Associative classifier
Minimal transversals
Hypergraphs
Microarray data

ABSTRACT

We investigate in this paper the problem of mining disjunctive emerging patterns in high-dimensional biomedical datasets. Disjunctive emerging patterns are sets of features that are very frequent among samples of a target class, cases in a case–control study, for example, and are very rare among all other samples. We, for the very first time, demonstrate that this problem can be solved using minimal transversals in a hypergraph. We propose a new divide-and-conquer algorithm that enables us to efficiently compute disjunctive emerging patterns in parallel and distributed environments. We conducted experiments using real-world microarray gene expression datasets to assess the performance of our approach. Our results show that our approach is more efficient than the state-of-the-art solution available in the literature. In this sense, we contribute to the area of bioinformatics and data mining by providing another useful alternative to identify patterns distinguishing samples with different class labels, such as those in case–control studies, for example.

1. Introduction

One of the challenges that analysts face when dealing with biomedical datasets relates to finding interesting patterns distinguishing samples in different classes in a case–control study. The task in these studies is to find patterns that are strongly correlated to samples in one target class, for instance case, and not correlated to samples belonging to the other class (controls). *Emerging patterns* are very suitable for this task. These patterns were introduced by Dong and Li [9] as an adaptation to the frequent itemset mining problem [1]. Emerging patterns are sets of features that are very frequent among samples of a target class and infrequent among samples of any other class.

The greatest advantage of these methods over other statistical techniques is their readability; since they are a collection of features that happen frequently in a group of samples, analysts and domain experts can easily read and interpret them. There are several examples of successful applications of both emerging patterns and frequent itemsets in biomedical studies such as, the works of Yeoh et al. [31], Creighton and Hanash [8], Gyenesei et al. [12] and Ramakrishnan and Zaki [25].

Even though these patterns have been successfully applied in bioinformatics, they still have some flaws if they are considered in exactly the same way as they were first introduced. First, they are mostly conjunctive, meaning that all features in a pattern have to occur together in a large number of samples to be considered frequent. The problem with conjunctions is that they do not account for heterogeneity of samples. In other words, a disease

* Corresponding author at: Centre for Bioinformatics, Biomarker Discovery and Information-Based Medicine, The University of Newcastle, University Drive, Callaghan, NSW 2308, Australia. Tel.: +61 2 4042 0510, fax: +61 2 4042 0037.
E-mail addresses: renato.vimieiro@newcastle.edu.au (R. Vimieiro), Pablo.Moscato@newcastle.edu.au (P. Moscato).

0306-4379/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.09.001

may be caused by the dysregulation of different molecular pathways and, in this case, part of samples labeled as cases may be affected by the dysregulation of genes in one pathway while the rest of the samples may be affected by the dysregulation of genes in a different pathway. Thus, there may not be a conjunction of genes that are common to all cases, but there could rather be a disjunction of genes that describe the whole set of cases. For instance, a disease could be associated to genes in one pathway or genes in another pathway.

The second problem is rather computational. The first algorithm was designed to mine data with a different nature; data where the number of samples was much bigger than the number of features. Biomedical datasets have a different format; they in general have many more features than samples.

There exist solutions for both finding disjunctive patterns [30,23] and efficient mining conjunctive patterns in biomedical data [24,7,15,19]. And there are even solutions for efficiently mining general Boolean patterns, rather than only conjunctions, in biomedical data [16,22].

We propose in this paper a new method to efficiently find disjunctive emerging patterns in biomedical data. Our approach is based on hypergraphs. For the first time, we model the problem of identifying disjunctive emerging patterns as minimal transversal enumeration problem using hypergraphs. Our approach allows us to use samples rather than features to find patterns. In this sense, this is the first algorithm to take full advantage of the relative scarcity of samples in biomedical datasets.

By showing that this problem can be broken into smaller subproblems and mapped onto the classical problem of enumerating *minimal transversals*, we allow it to be easily solved in parallel and/or distributed environments. Our experiments demonstrate the effectiveness of such an approach. Thus, we believe that our algorithm is indeed an important tool for bioinformaticians, and other data scientists alike.

We conducted several experiments to assess the computational performance of our method using real-world datasets obtained from the Gene Expression Omnibus website. Our experiments show that our method can efficiently find interesting emerging patterns in high-dimensional biomedical data.

This paper is organized in the following manner. In the next section we formally define the concepts that we briefly introduced in this section. We define disjunctive emerging patterns, some important properties of these patterns and provide more details on the problem that we address. In Section 3 we introduce our main algorithm. As we discuss later in this paper, we model the problem of mining disjunctive emerging patterns as a minimal transversal enumeration problem in order to benefit from the vast literature in this topic. In this case, we do not propose a new algorithm for enumerating minimal transversals in a hypergraph, but rather use efficient solutions in this area to solve our problem. In Section 4 we briefly discuss the algorithms that we identified in the literature and used to assess the performance of our method in Section 5. We finish the paper with some final remarks in Section 6.

2. Disjunctive emerging patterns

Let $A_1, A_2, \ldots, A_n$ be a set of n categorical attributes on a dataset. Let F denote the set of features in a dataset; the union of all possible values for the attributes. Let S be a set of samples, where each sample contains *exactly one* value for each attribute in the dataset. For a given sample $s \in S$, we define $f(s)$ to be the set of features associated with s, in other words, $f(s) = \bigcup A_i(s)$, where $A_i(s)$ is the value that s has for attribute A_i. We restrict our problem to the case of binary classes, which we call positive and negative. In this case, the class label defines a partition on the set of samples S^+ and S^-, where S^+ is the subset of samples with positive class label and S^- is the subset of samples with negative class label.

A *disjunctive emerging pattern* (DEP) is a subset of features, $X \subseteq F$, fulfilling the following constraints [16]:

1. X includes at least one value from each attribute.
2. X occurs (evaluates to true) in *at least* α samples with positive class label.
3. X occurs (evaluates to true) in *at most* β samples with negative class label.

We say that a set of features X occurs in a sample $s \in S$ if $f(s) \subseteq X$. Then, X is a disjunctive emerging pattern if and only if $g^+(X) = |\{s \in S^+ | f(s) \subseteq X\}| \geq \alpha$ and $g^-(X) = |\{s \in S^- | f(s) \subseteq X\}| \leq \beta$.

A DEP X is *maximal* if there is no proper superset of X that is also a DEP. On the other hand, we say that X is a *jumping disjunctive emerging pattern* if it occurs in the positive class (i.e. in at least one sample from positive class) and does not occur at all in the negative class. Thus, X is a jumping DEP if X is a DEP for $\alpha = 1$ and $\beta = 0$.

The number of DEPs in a dataset can be extremely large. This problem, as discussed by Vimieiro and Mascato [30] and Vimieiro [29], is not exclusive of DEPs and it does occur for other types of frequent patterns, like traditional conjunctive frequent patterns or disjunctive patterns. Then, it is desirable to have a concise set that represents the entire set of DEPs.

For traditional conjunctive frequent patterns, this was achieved by extracting maximal frequent patterns. The anti-monotocity of the frequency of conjunctive patterns guarantees that every subset of a maximal frequent pattern is still a frequent pattern, and, therefore, it is viable and sound to mine only these patterns. Loekito and Bailey [16] also attempted to use maximal sets to compactly represent the whole set of DEPs. However, although the set of maximal itemsets is a valid compact representation for frequent itemsets, the same cannot be said about the set of maximal DEPs.

In fact, rather than an erroneous representation, a compact representation using only maximal DEPs is an incomplete one. Maximal sets define the positive border of the search space [18]. In our context, it means that, if we are searching for patterns level-wise, once we find all maximal DEPs, we can stop searching for more patterns, because no proper superset of a maximal DEP can still be valid. In other words, maximal DEPs are necessary to