Expert Systems with Applications 41 (2014) 5105–5114

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Novel techniques and an efficient algorithm for closed pattern mining

András Király [a,*], Asta Laiho [b], János Abonyi [a], Attila Gyenesei [b]

[a] *University of Pannonia, Department of Process Engineering, P.O. Box 158, Veszpreém H-8200, Hungary*
[b] *The Finnish Microarray and Sequencing Centre, Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, Tykistökatu 6A, 20520 Turku, Finland*

ARTICLE INFO

Keywords:
Biclustering
Closed frequent itemset mining
Clustering visualization
Data mining algorithm
Pattern detection

ABSTRACT

In this paper we show that frequent closed itemset mining and biclustering, the two most prominent application fields in pattern discovery, can be reduced to the same problem when dealing with binary (0–1) data. FCPMiner, a new powerful pattern mining method, is then introduced to mine such data efficiently. The uniqueness of the proposed method is its extendibility to non-binary data. The mining method is coupled with a novel visualization technique and a pattern aggregation method to detect the most meaningful, non-overlapping patterns. The proposed methods are rigorously tested on both synthetic and real data sets.

1. Introduction

Very large datasets have lately become increasingly common in many application areas, making it impossible to inspect data manually when looking for interesting patterns and knowledge (Tan, Steinbach, & Kumar, 2006). Data mining as a field aims at developing computational methods and tools that can be used for automating the knowledge extraction for the aid of decision making or for understanding general trends within data collections (Kantardzic, 2002). Pattern discovery, an interesting subfield of data mining, is motivated by the huge amounts of electronic data that many organizations produce on their functions or collect during experiments within various research fields. For instance, supermarkets store electronic copies of millions of receipts, banks and credit card companies maintain extensive transaction histories and biomedical and bioinformatics research groups collect data from various biological experiments. The goal of pattern discovery in general is to analyze these large datasets to identify informative patterns and motifs (Berry & Linoff, 1997).

Frequent pattern discovery is an actively researched data mining technique with a wide range of applications (Han, Cheng, Xin, & Yan, 2007). Market basket data analysis in web-shops, web search analysis, frequent symptom set mining in health-care or proper product placement in supermarkets are all well-known association rule mining techniques, where pattern matching is used to predict the behavior of customers or patients. Using adjacency matrices, strongly connected components can be revealed in social networks and by the help of collaborative filtering automated suggestions can be performed for users taking into account the duality between users and items. In the field of genetics, the huge amount of data from gene expression analysis can be handled by efficient pattern mining algorithms to uncover local motifs and interesting genetic pathways which are not apparent otherwise.

Frequent itemset mining methods usually generates a huge amount of patterns or association rules so the result of the data mining is not directly applicable. One possible solution to get a compact and informative set of itemsets is closed itemset mining (Agrawal, Imieliński, & Swami, 1993; Brin, Motwani, Ullman, & Tsur, 1997). During the last decade many frequent closed itemset mining methods have been proposed in the literature. The problem was firstly introduced by Pasquier et al. in Pasquier, Bastide, Taouil, and Lakhal (1999) together with the first algorithm called A-Close for mining closed itemsets. Other closed itemset mining algorithms include CLOSET (Pei, Han, & Mao, 2000), CHARM (Zaki & Hsiao, 1999), CLOSET+ (Wang, Han, & Pei, 2003), FPClose (Grahne & Zhu, 2003), AFOPT (Liu, Lu, Lou, & Yu, 2003), DCI_Closed (Lucchese, Orlando, & Perego, 2006), DBV-Miner (Rodríguez-González, Martínez-Trinidad, Carrasco-Ochoa, & Ruiz-Shulcloper, 2013), DBV-Miner (Vo, Hong, & Le, 2012) and ClaSP (Gomariz, Campos, Marin, & Goethals, 2013) just to mention the most applied ones. Very recent publications are presented in Zhou, Cule, and Goethals (2013), Riondato and Upfal (2013), Cule, Goethals, and Hendrickx (2013) and methods for approximation of true frequent itemsets can be found in Riondato and Vandin (2014) and Riondato and Upfal (2013). For comprehensive reviews about the efficient algorithms, see Fimi (2003), Fimi (2004) and Duneja and Sachan (2012).

[*] Corresponding author. Tel.: +36 88624770; fax: +36 88623171.
E-mail address: kiralya@fmt.uni-pannon.hu (A. Király).

http://dx.doi.org/10.1016/j.eswa.2014.02.029
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Based on the critical analysis of literature we think beside of contributions aiming performance improvements, there is a place for a novel approach that is not only accurate and effective but gives more insight to the hidden structure of the itemsets. Therefore we are looking for a novel method that is intuitive, supports visualization and allows further aggregation of the found patterns.

Our main idea is that the problem of finding closed frequent itemsets can be considered as mining biclusters in binary data, so the favorable properties of biclustering representation will ensure the required interpretability.

The fields of frequent closed itemset mining and biclustering were developed independently. Biclustering has been introduced to complement and expand the capabilities of the standard clustering methods by allowing objects to belong to multiple or none of the resulting clusters purely based on their similarities. This property makes biclustering a powerful approach especially when it is applied to data with a large number of objects. During recent years, many biclustering algorithms have been developed especially for the analysis of gene expression data. The concept of biclustering was first introduced in Hartigan (1972), and applied to gene expression data by Cheng and Church (2000). Many other such algorithms have been published since including the most refereed ones including BiMAX (Prelić et al., 2006), QUBIC (Li, Ma, Tang, Paterson, & Xu, 2009), BiBit (Rodriguez-Baena, Perez-Pulido, & AguilarRuiz, 2011), Signature Algorithm (Ihmels, Bergmann, & Barkai, 2004), xMotif (Murali & Kasif, 2003), OPSM (Ben-Dor, Chor, Karp, & Yakhini, 2003) and many others (e.g. Abdullah & Hussain (2006), Uitert, Meuleman, & Wessels (2008) and Cheng, Law, & Siu (2013)). For comprehensive reviews, see Busygin, Prokopyev, and Pardalos (2008), Kriegel, Kröger, and Zimek (2009), Bozdağ, Kumar, and Çatalyürek (2010), Eren, Deveci, Küçüktunç, and Çatalyürek (2013) and Freitas, Ayadi, Elloumi, Oliveira, and Hao (2013). As for biclustering has been mainly used for gene expression data analysis with the aim of discovering gene expression patterns or so-called biclusters (Madeira & Oliveira, 2004). In the later case, biclustering instead of using original real valued data, the input is discretized in order to reduce the dimensionality and enable reasonable processing times. Typically the data is transformed into a binary matrix containing only 0 and 1 values prior to applying a mining algorithm.

One of our main goals is to show that frequent closed itemset mining and biclustering of binary data can be transformed to the same problem and therefore, all existing methods for mining such patterns in binary data can in fact be applied to both fields. This finding might also help researchers to identify new research directions. Indeed, as a first step in this direction we will extend the problem of mining patterns in binary data by introducing a novel and efficient method that is able to discover previously hidden patterns with more than two value categories.

Since data mining often produces a large number of small frequent and partially overlapping patterns, this causes a considerable challenge for the result interpretation. Therefore, we introduce a novel visualization that rearranges the original data matrix based on the discovered closed patterns and a pattern aggregation method allowing the rapid identification of the most meaningful pattern clusters.

The contributions of this work will be discussed in the following subsections as novel methods and algorithms:

- After introducing and defining the problems of biclustering and frequent closed itemset mining on binary data, we show that they can be reduced to the same problem (Sections 2.1–2.3). We prove this both theoretically (Section 2.3) and experimentally (Section 4) on various synthetic and real data sets.
- We extend the problem of mining frequent closed patterns (FCP) to more than two value categories. This is especially important when FCP mining methods are applied to big biological, such as gene expression data that are commonly obtained nowadays by microarray and next-generation sequencing instruments (see Section 3.1).
- We propose a novel algorithm, called FCPMiner to mine FCPs efficiently. We rigorously test FCPMiner on various real and synthetic data sets and show that FCPMiner is a powerful method for mining FCPs (see Section 3.1.1).
- We introduce a novel visualization as well as a pattern aggregation method to enable the quick identification of the most relevant FCPs (see Sections 3.2 and 3.3).
- We implemented our algorithms using the Java programming language that can be run on any Operation System (Windows, Linux, Mac OS). The implemented methods are freely available on the project website.3

2. Problem formulation

2.1. Biclustering

In this paper we follow the formulation given in Prelić et al. (2006) to define the problem of mining biclusters in gene expression data. According to common practice of the field, bicluster mining is restricted to a binary matrix, i.e. gene expression values are transformed to 1 (expressed) or 0 (not expressed) using an expression cutoff (Li et al., 2009; Prelić et al., 2006). Let $E \in \{0,1\}^{n \times m}$ be an expression matrix, where E represents the set of m experiments for n genes. A cell e_{ij} contains 1 whenever gene i is expressed in condition j and 0 otherwise. A bicluster (G,C) corresponds to a subset of genes $G \subseteq \{1,\dots,n\}$ that jointly responds a subset of samples $C \subseteq \{1,\dots,m\}$. Therefore, the bicluster (G,C) is a submatrix of E in which all elements are equal to 1 (biclusters in a small data set are depicted in Fig. 1 by bold numbers). Using the above definition, every cell e_{ij} having only non-zero values represents a bicluster. However, such patterns are usually redundant as they are entirely contained by other patterns. Thus, the definition of *inclusion-maximal bicluster* (*IMB*) was introduced to discover all biclusters not entirely contained by any other cluster (Prelić et al., 2006): the pair $(G,C) \in 2^{\{1,\dots,n\}} \times 2^{\{1,\dots,m\}}$ is an *IMB*, if and only if $\forall i \in G, j \in C : e_{ij} = 1$ and $\nexists (G',C') \in 2^{\{1,\dots,n\}} \times 2^{\{1,\dots,m\}}$ where $\forall i' \in G', j' \in C' : e_{i'j'} = 1$ and $G \subseteq G' \wedge C \subseteq C' \wedge (G',C') \neq (G,C)$.

2.2. Frequent closed itemset mining

One of the earliest and most important concepts in data mining is mining frequent itemsets in large transactional datasets (Lucchese, Orlando, & Perego, 2010). Such a dataset can be considered as a matrix with transactions as rows and items as columns. If an item appears in a transaction it is denoted by 1, otherwise by 0. The general goal of frequent itemset mining is to identify all itemsets that contain at least as many transactions as required, referred to as *minimum support threshold*, *min_sup*. By definition, all subsets of a frequent itemset are frequent. Therefore, it is also important to provide a minimal representation of all frequent itemsets without losing their support information. Such itemsets are called *frequent closed itemsets* and can be defined as follows. Let $\sigma(x) = |\{t_i : x \subseteq t_i, t_i \in \mathcal{T}\}|$ denote the support count of itemset x. An itemset x is *closed* if none of its immediate supersets has exactly the same support count as x. Oppositely, the itemset x is not closed if at least one of its immediate supersets has the same support count as x. Obviously, $x, y : \sigma(x) \geqslant$