Information Sciences 420 (2017) 1–15

Contents lists available at ScienceDirect

Information Sciences

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

Combining Apriori heuristic and bio-inspired algorithms for solving the frequent itemsets mining problem

Youcef Djenouri, Marco Comuzzi*

Ulsan National Institute of Science and Technology (UNIST), 50 UNIST-gil, Ulju-gun, Ulsan 44919, Republic of Korea

ARTICLE INFO

Article history:
Received 25 October 2016
Revised 7 August 2017
Accepted 11 August 2017
Available online 14 August 2017

Keywords:
Frequent itemsets mining
Apriori heuristic
Genetic algorithm
Particle swarm optimization

ABSTRACT

Exact approaches to Frequent Itemsets Mining (FIM) are characterised by poor runtime performance when dealing with large database instances. Several FIM bio-inspired approaches have been proposed to overcome this issue. These are considerably more efficient from the point of view of runtime performance, but they still yield poor quality solutions. The quality of the solution, i.e., the number of frequent itemsets discovered, can be increased by improving the randomised search of the solutions space considering intrinsic features of the FIM problem. This paper proposes a new framework for FIM bio-inspired approaches that considers the recursive property of frequent itemsets, i.e., the same feature exploited by the Apriori exact heuristic, in the search of the solution space. We define two new approaches to FIM, namely GA-Apriori and PSO-Apriori, based on the proposed framework, which use genetic algorithms and particle swarm optimisation, respectively. Extensive experiments on synthetic and real database instances show that the proposed approaches outperform other bio-inspired ones in terms of runtime performance. The results also reveal that the performance of PSO-Apriori is comparable to the one of exact approaches Apriori and FPGrowth in respect of the quality of solutions found. We also show that PSO-Apriori outperforms the recently developed BATFIM algorithm when dealing with very large database instances.

1. Introduction

Frequent Itemsets Mining (FIM) aims at extracting frequent items highly correlated from a transactional database. The FIM problem is defined as follows. Let T be a set of M transactions $T = \{t_1, t_2, \dots, t_M\}$, representing a transactional database, and I be a set of N different items (or attributes) $\{i_1, i_2, \dots, i_N\}$. An itemset X is a set of items, i.e., $X \subset I$. The support of an itemset $Sup(X)$ is the number of transactions that contains X divided by M. An itemset X is *frequent* if its support is no less than *MinSup* [1], where *MinSup* is a threshold chosen by the user.

Exact approaches to FIM, such as Apriori [1] and FPGrowth [10], find all frequent itemsets in a database. Because of the need for multiple scans of an entire transactional database, the performance of exact algorithms tends to quickly degrade with the size of the database and become unacceptable for large database instances, such as data generated from social networks [14] or large bioinformatics datasets [28]. For instance, [5] reports that FIM exact approaches crashed or did not complete even after 10 days of runtime while processing the WebDocs[1] database instance.

* Corresponding author.
E-mail addresses: ydjenouri@unist.ac.kr (Y. Djenouri), mcomuzzi@unist.ac.kr (M. Comuzzi).

[1] This database is also used in the experimental evaluation of this paper.

http://dx.doi.org/10.1016/j.ins.2017.08.043
0020-0255/© 2017 Elsevier Inc. All rights reserved.

To overcome the performance issue of exact approaches, several approaches to FIM have been proposed that use bio-inspired techniques, such as genetic algorithms [6,22] or swarm intelligence [7,19]. These tend to perform in reasonable time, but they do not find all possible frequent itemsets in a database. In other words, the quality of the solution obtained using bio-inspired approaches is lower than the optimal quality achieved by exact approaches, which discover all possible frequent itemsets. Therefore, particularly in the context of very large instances, there is a continuous challenge in the data mining community to propose bio-inspired FIM approaches that return higher quality solutions, while maintaining reasonable runtime.

The quality of the solutions in bio-inspired FIM approaches relies on the way in which the randomised search of the itemsets space is performed. We argue that FIM bio-inspired approaches in the literature fail to consider the intrinsic properties of the FIM solution space to improve such a search. The most noticeable of these properties is that frequent itemsets are *recursive*, i.e., if an itemsets of size k is frequent, then all of its sub-itemsets of size $s = 1, \ldots, k-1$ are also frequent. This feature is at the cornerstone of the Apriori exact heuristic, but it is not usually exploited by FIM bio-inspired approaches. Genetic algorithms-based FIM approaches [22], for instance, differ in the way in which individual population elements are coded or local solutions are combined to obtain new ones, but they all consider populations of itemsets of various size at each iteration. Similarly, particle swarm optimisation approaches, such as [19], randomly initialise a set of particles identifying itemsets of different size in the solution space and then adjust their velocity to update their positions.

This paper proposes a new framework for designing FIM bio-inspired approachers that exploits the *recursive* property of frequent itemsets in a database. In our framework, we propose to start exploring the FIM search space with a randomly initialised population of size $k = 1$. Then, at a generic iteration k, the intensification and diversification operators of the chosen bio-inspired metaheuristic are defined to produce only individuals, i.e., itemsets, of size k starting from a population of itemsets of size $k-1$. In this way, the probability that itemsets in the newly generated population are frequent highly increases since, instead of generating random individuals of different sizes, individuals of size k are generated only by recombining the features of frequent itemsets of size $k-1$. Moreover, the number of search iterations is capped to K, that is, the number of itemsets (size) of the largest transaction in a database.

Based on our framework, we propose two new bio-inspired FIM algorithms, i.e., GA-Apriori and PSO-Apriori, that use genetic algorithms and particle swarm optimisation, respectively. These two new algorithms improve the existing GA-FIM [6] and PSO-FIM [19] algorithms by defining search space intensification and diversification operators, i.e., crossover and mutation for GA-FIM and particle positioning and velocity for PSO-FIM, that take into account the recursive property of frequent itemsets.

Extensive experiments have been run on synthetic and real data instances to validate the performance of GA-Apriori and PSO-Apriori. The results show that they outperform the existing GAFIM and PSOFIM algorithms in terms of both the number of frequent itemsets discovered and runtime performance. Moreover, they outperform exact algorithms, i.e., Apriori and FPGrowth, in terms of runtime performance. The results also reveal that PSO-Apriori is competitive compared to the exact approaches in respect of the quality of solutions found. The last experiments shows that PSO-Apriori outperforms the recently developed BATFIM [12] algorithm, based on bat swarms intelligence, when dealing with very large data instances.

This paper advances the state of the art of bio-inspired FIM approaches by proposing two new algorithms that improve the quality of the solution found, i.e., the number of frequent itemsets discovered, without worsening the runtime performance. Moreover, the proposed framework paves the way for the development of new FIM approaches using alternatives bio-inspired metaheuristics for which the intensification and diversification operators take into account the recursive property of the FIM solution space.

The remainder of the paper is organized as follows: Section 2 reviews related work on solutions to the FIM problem and provides required background knowledge on the GAFIM and PSOFIM algorithms and the Apriori heuristic. The framework and, in particular, the GA-Apriori and PSO-Apriori algorithms are presented in Section 3. The experimental evaluation is described in Section 4, and finally, Section 5 draws the conclusions.

2. Background and related work

This section briefly introduces the Apriori heuristic, which exploits the recursive property of frequent itemsets to devise an exact search algorithms. Then, we review related work in the area of bio-inspired FIM approaches and, finally, we discuss in detail the existing algorithms GA-FIM and PSO-FIM, which are extended by the algorithms proposed later in this paper.

2.1. Apriori heuristic

The Apriori heuristic considers that an itemset of size k is frequent if and only if all its subsets are frequent. Thus, at each iteration k, the candidates itemsets of size k are generated by joining two frequent itemsets of size $k-1$. This process is repeated until the candidate itemsets of length k is empty.

Let us consider an example database containing 5 transactions $\{t_1$: $\{a, b\}$, t_2: $\{b, c, d\}$, t_3: $\{a, b, c\}$, t_4: $\{e\}$, t_5: $\{c, d, e\}\}$. Fig. 1 illustrates the results of applying the Apriori algorithm considering *MinSup* equal to 40%. The database is first scanned to calculate the support of each candidate itemset of size 1, i.e., candidate itemsets containing only one item. The frequent itemsets of size 1 are then extracted (see L1 in Fig. 1). In the example, all candidates itemsets of size 1 are frequent because their supports are greater than 0.4. In the second iteration, the candidate itemsets of size 2 (C2) are extracted by joining the