Expert Systems With Applications 93 (2018) 62–71

Contents lists available at ScienceDirect

Expert Systems With Applications

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

OOIMASP: Origin based association rule mining with order independent mostly associated sequential patterns

Deepak Yadav, C. Ravindranath Chowdary*

Department of Computer Science and Engineering, Indian Institute of Technology (BHU) Varanasi, 221005, India

ARTICLE INFO

Article history:
Received 5 July 2017
Revised 5 October 2017
Accepted 5 October 2017
Available online 7 October 2017

Keywords:
Association rule mining
Mostly associated sequential patterns
Unbiased support
Unbiased confidence

ABSTRACT

Efficient mining of association rules on a transaction dataset is an interesting and a challenging problem. The state-of-the-art MASP algorithm is dependent on the order of items in the transaction. We propose OOIMASP algorithm, which has two novel properties- 1) order independence and 2) it takes into consideration the origin of items to calculate unbiased support and unbiased confidence values. Order dependence is one of the drawbacks of MASP. OOIMASP addresses this issue by rearranging the items in transactions using a greedy frequency based approach. We compare the performance of our system with MASP on five synthetic data sets and three public data sets. The results show that our proposed approach outperforms the MASP in both the comparison metrics, i.e., the number of association rules generated and the length of the longest association rule. Both these metrics are important to evaluate the performance of an algorithm. On an average, OOIMASP algorithm generates 632% longer rules and 457% more association rules than MASP algorithm. The disadvantage of the proposed algorithm is, it requires more computational resources in terms of time, approximately 5 times more than MASP. We claim that the extra information extracted using our method compensates for the increase in time complexity as compared to MASP. The proposed method produces multiple trees which can be very useful in the visual analysis of data.

1. Introduction

Association rule mining can be used to find interesting patterns in the transaction databases based on individual and conditional frequencies. In the conventional approach, two phases are involved in generating rules. In the first phase, all the itemsets are generated and in-frequent itemsets are pruned. In the second phase, rules are derived from those frequent itemsets. An association rule, e.g., {bread, milk} ⇒ {butter} in market basket analysis means if a user purchases bread and milk together, it is highly likely that she will also purchase butter. Apart from market basket analysis, association rule mining is useful in intrusion detection (Mabu, Chen, Lu, Shimada, & Hirasawa, 2011), medical diagnosis (Sharma et al., 2015), protein sequences (Gupta, Mangal, Tiwari, & Mitra, 2006), Customer Relationship Management (CRM) (Song, kyeong Kim, & Kim, 2001) and many other applications.

Soysal (2015) proposed a novel approach to extract Mostly Associated Sequential Patterns (MASPs) using less computational resources in terms of time and memory while generating a long sequence of patterns that have the highest co-occurrence. This approach may produce different outcomes if we change the order of items in transactions. Fig. 1 shows this issue. In Fig. 1, the MASP tree changes with the change in the order of items in transactions for threshold support 0.50. We propose an approach which is order independent. An association rule of the form $A \Rightarrow B$ must satisfy the threshold support and threshold confidence. To compute support and confidence, it is required to traverse the complete transaction database. What if an item appears for the first time in the ith transaction? It is biased to take the entire dataset for calculating support and confidence for the rules containing that particular item. So, to generate all the rules containing a particular item x, it is reasonable to ignore all transactions (for calculating support and confidence) that come before the transaction in which that particular item appears for the first time. Taking into account the origin of the item is our second contribution.

Our contributions:

1) Proposed an algorithm to generate order independent Mostly Associated Sequential Patterns (modified the state-of-the-art algorithm (Soysal, 2015))
2) Considered the origin of items for computing unbiased support and unbiased confidence

* Corresponding author.
E-mail addresses: deepak.yadav.cse14@iitbhu.ac.in (D. Yadav), rchowdary.cse@iitbhu.ac.in (C.R. Chowdary).

https://doi.org/10.1016/j.eswa.2017.10.015
0957-4174/

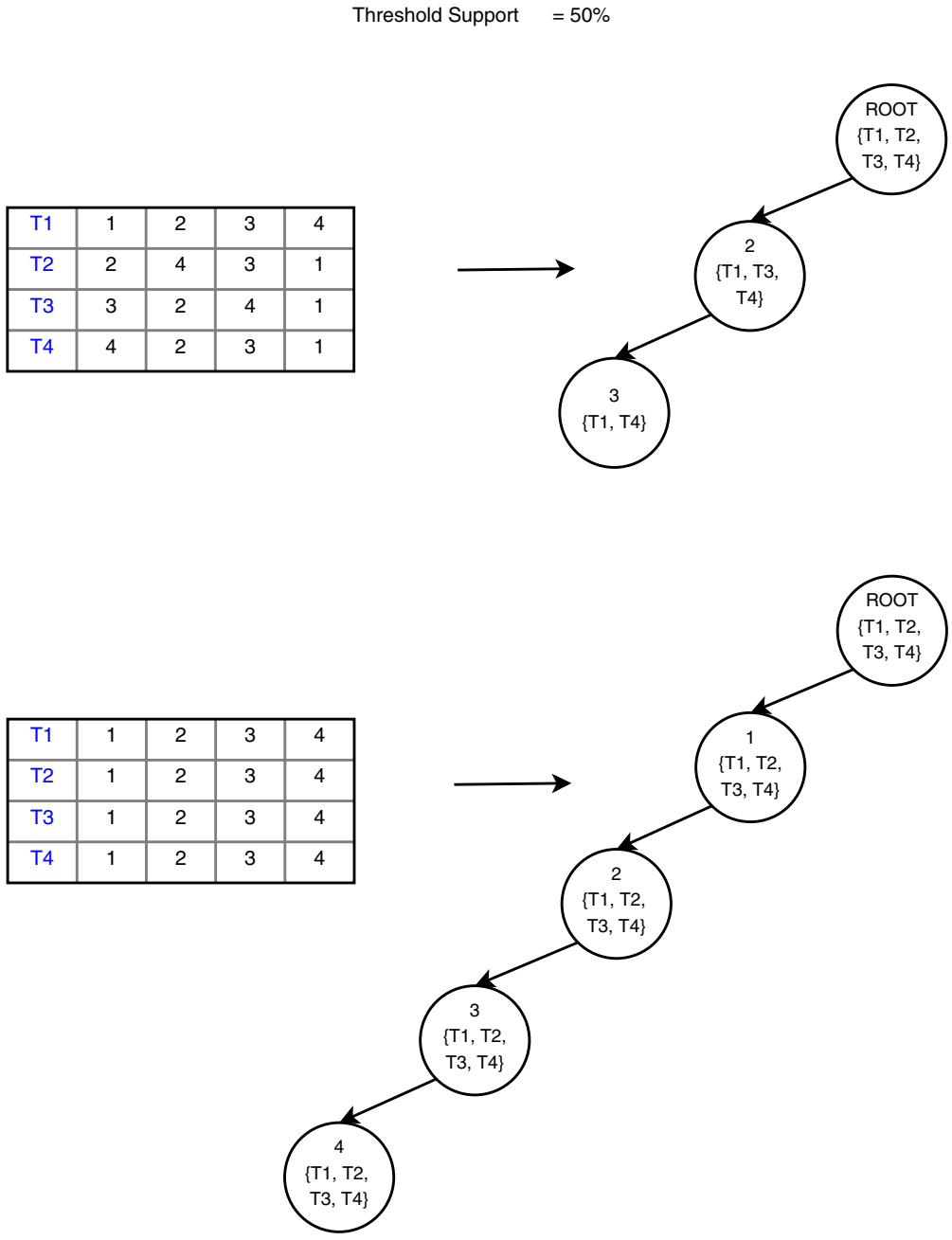

Fig. 1. Different *MASP* trees for the same dataset.

The rest of the paper is organized as follows: Section 2 discusses the literature on association rule mining. Terms and definitions are introduced in Sections 3 and 4 discusses the proposed algorithms. Experimental setup and results are discussed in Section 5 followed by Conclusions in Section 6.

2. Related work

Agrawal and Srikant (1994) proposed apriori algorithm for finding association rules in large databases of the sales transaction. Apriori algorithm produces association rules in two phases. In the first phase, all the itemsets are generated, and infrequent itemsets are pruned. In the second phase, rules are derived from those frequent itemsets. This algorithm first finds frequent itemsets of length one then frequent itemsets of length two using frequent itemsets of length one and so on until the generation of all the frequent itemsets. This algorithm performs better than the previously known fundamental algorithms AIS (Agrawal, Imieliński, & Swami, 1993), SETM (Houtsma & Swami, 1993). Fukuda, Morimoto, Morishita, and Tokuyama (1996) proposed an approach to find two-dimensional association rules. A state in this scenario is of the form $((X, Y) \in P) \Rightarrow (Z = z)$ where X and Y are numeric attributes, P is a subspace of 2-D plane, and Z is a boolean attribute, i.e., z can be either true or false. E.g. $(Age \in [30, 50] \wedge Balance \in [10^5, 10^6]) \Rightarrow (CardLoan = \text{yes})$. It means if a bank user age and balance lies in the given subspace it is very likely that she will use card loan. This approach works for specific types of structured data.

Feldman, Aumann, Amir, Zilberstein, and Kloesgen (1997) introduced the notion of maximal association rules. These are the rules extracted from frequent maximal itemsets. Frequent maximal itemsets are those itemsets which appear just once among all the transactions. It is useful in finding association rules con-