Objectively Evaluating Interestingness Measures for Frequent Itemset Mining

Albrecht Zimmermann

KU Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium
albrecht.zimmermann@cs.kuleuven.be

Abstract. Itemset mining approaches, while having been studied for more than 15 years, have been evaluated only on a handful of data sets. In particular, they have never been evaluated on data sets for which the ground truth was known. Thus, it is currently unknown whether itemset mining techniques actually recover underlying patterns. Since the weakness of the algorithmically attractive support/confidence framework became apparent early on, a number of interestingness measures have been proposed. Their utility, however, has not been evaluated, except for attempts to establish congruence with expert opinions. Using an extension of the Quest generator proposed in the original itemset mining paper, we propose to evaluate these measures objectively for the first time, showing how many non-relevant patterns slip through the cracks.

1 Introduction

Frequent itemset mining (FIM) was introduced almost twenty years ago [1] and the framework has proven to be very successful. It not only spawned related approaches to mining patterns in sequentially, tree, and graph-structured data, but due to its relative simplicity it has been extended beyond the mining of supermarket baskets towards general correlation discovery between attribute value pairs, discovery of co-expressed genes, and classification rules, etc.

The original framework used frequency of itemsets in the data (support) as a significance criterion – often occurring itemsets are assumed not to be chance occurrences – and conditional probability of the right-hand side of association rules (confidence) as a correlation criterion. This framework has clear weaknesses and other interestingness measures have been proposed in the years since the seminal paper was published [18], as well as several condensed representations [13,5,10] that attempt to remove redundant information from the result set.

While each of these measures and condensed representations is well-motivated, there is as of yet no consensus about how effectively existing correlations are in fact discovered. A prime reason for this can be seen in the difficulty of evaluating the quality of data mining results. In classification or regression tasks, there is a clearly defined target value, often objectively measured or derived from expert labeling *a priori* to the mining/modeling process, that results can be compared to to assess the goodness of fit. In clustering research, the problem is somewhat more

J. Li et al. (Eds.): PAKDD 2013 Workshops, LNAI 7867, pp. 354–366, 2013.

pronounced but clusters can be evaluated w.r.t. intra-cluster similarity and inter-clusters dissimilarity, knowledge about predefined groups might be available, e.g. by equating them with underlying classes, and last but not least there exist generators for artificial data [15]. In FIM, in contrast, while the seminal paper introduced a data generator as well, that data generator has been used only for efficiency estimations and fell furthermore into some disregard after Zheng *et al.* showed that the data it generated had characteristics that were not in line with real-life data sets [22]. The current, rather small collection of benchmark sets, hosted at the FIMI repository [2], consists of data sets whose underlying patterns are unknown. As an alternative, patterns mined using different measures have been shown to human "domain experts" who were asked to assess their interestingness [8]. Given humans' tendency to see patterns where none occur, insights gained from this approach might be limited.

Interestingly enough, however, the Quest generator proposed by Agrawal *et al.* already includes everything needed to perform such assessments: it generates data by embedding source itemsets, making it possible to check mining results against a gold standard of predefined patterns. Other data generation methods proposed since [16,9,17,3] do not use clearly defined patterns and can therefore not be used for this kind of analysis.

The contribution of this work is that we repurpose the Quest generator accordingly and address this open questions for the first time:

– How effective are different interestingness measures in recovering embedded source itemsets?

In the next section, we introduce the basics of the FIM setting, and discuss different interestingness measures. In Section 3, we describe the parameters of the Quest generator and its data generation process. Equipped with this information, we can discuss related work in Section 4, placing our contribution into context and motivating it further. Following this, we report on an experimental evaluation of pattern recovery in Section 5, before we conclude in Section 6.

2 The FIM Setting

We employ the usual notations in that we assume a collection of *items* $\mathcal{I} = \{i_1, \ldots, i_N\}$, and call a set of items $I \subseteq \mathcal{I}$ an itemset, of size $|I|$. In the same manner, we refer to a transaction $t \subseteq \mathcal{I}$ of size $|t|$, and a data set $\mathcal{T} \subseteq 2^{\mathcal{I}}$, of size $|\mathcal{T}|$. An itemset I matches (or is supported by) a transaction t iff $I \subseteq t$, the support of I is $sup(I, \mathcal{T}) = |\{t \in \mathcal{T} \mid I \subseteq t\}|$, and its relative support or frequency $freq(I, \mathcal{T}) = \frac{sup(I,\mathcal{T})}{|\mathcal{T}|}$. The *confidence* of an association rule formed of two itemsets $X, Y \subset \mathcal{I}, X \cap T = \emptyset$ is calculated as $conf(X \Rightarrow Y, \mathcal{T}) = \frac{sup(X \cup Y,\mathcal{T})}{sup(X,\mathcal{T})}$. When the context makes it clear which data set is referred to, we drop $\mathcal{T}$ from the notation.