Objectively evaluating interestingness measures for frequent itemset mining

Albrecht Zimmermann
albrecht.zimmermann@cs.kuleuven.be

KU Leuven, Celestijnenlaan 200A, B-3001 Leuven, Belgium

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. As a result of this, 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 measure 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 since. Not only did it spawn related approaches to mining patterns in sequentially, tree, and graph-structured data, but due 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, to name a few.

The original framework used frequency of itemsets in the data (support) as a significance criterion – itemsets that occur often are assumed not to be chance occurrences – and conditional probability of the right-hand side of association rules (confidence) as a correlation criterion. Both of these measures have clear weaknesses, however, and a number of other interestingness measures have been proposed in the years since the seminal paper was published [2], as well as several condensed representations [3–6] 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 algorithmic results can be compared to to assess the goodness of fit. In clustering research, the problem is somewhat more 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 [7]. 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 [8]. The current, rather small collection of benchmark sets, hosted at the FIMI repository [9], 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 [10]. 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 [11–17] do not use clearly defined patterns and can therefore not be used for this kind of analysis. Furthermore, data can be generated using different combinations of the number of items and source itemsets in the data, allowing to simulate different data densities which would allow to do more comprehensive run time experiments. And of course, different combinations of transaction and itemset sizes can be used.

The contribution of this work is that we repurpose the Quest generator accordingly and address one of the 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 $I \subseteq t$, and 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.

2.1 Interestingness measures

The support/confidence framework has at least one major drawback in that it ignores prior probabilities. Assume, for instance, two items i_1, i_2 with $freq(i_1) = 0.6$, $freq(i_2) = 0.8$. While $freq(i_1, i_2) = 0.48$ would often denote the itemset as a high-frequency itemset, it is in fact exactly what would be expected given independence of the two items. Similarly, $conf(i_1 \Rightarrow i_2) = 0.8$, while clearly a high confidence value, would also indicate independence when compared to the prior frequency of i_2. Therefore, numerous other measures have been proposed to address these shortcoming [18].

Most of them have been proposed for assessing the quality of association rules, meaning that they relate two binary variables. Generally speaking, it is possible to use such measures more generally to assess the quality of itemsets in the following way. Given an interestingness measure $m : \mathcal{I} \times \mathcal{I} \mapsto \mathbb{R}$, itemset I, we can take the minimal value over all possible association rules with a single item in the right-hand side (RHS): $\min_{i \in I}\{m(I \setminus i \Rightarrow i\})$. This is the approach taken in the FIM implementations of Christian Borgelt[1] and since we employ those in our experiments, we evaluated the included additional measures as well.

Our primary aim, however, is to test the recovery of itemsets, the precursors to association rules, and we therefore focus on measures that have been proposed to mine interesting *itemsets*. To make it easier to discuss those more sophisticated measures, we associate each itemset with a function $I : \mathcal{T} \mapsto \{0, 1\}$, with $I(t) = 1$ iff $I \subseteq t$ and $I(t) = 0$ otherwise, which allows us to define an equivalence relation based on a collection of itemsets $\{I_1, \ldots, I_k\}$:

$$\sim_{\{I_1,\ldots,I_k\}} = \{(t_1, t_2) \in \mathcal{T} \times \mathcal{T} \mid \forall I_i, I_j : I_i(t_1) = I_j(t_2)\}$$

Using this equivalence relation, the *partition* or quotient set of $\mathcal{T}$ over $\{I_1, \ldots, I_k\}$ is defined as:

$$\mathcal{T}/\sim_{\{I_1,\ldots,I_k\}} = \bigcup_{t \in \mathcal{T}} \{a \in \mathcal{T} \mid a \sim_{\{I_1,\ldots,I_k\}} t\}$$

We label each block $b \in \mathcal{T}/\sim_{\{I_1,\ldots,I_k\}}$ with a subscript denoting what the different itemsets evaluate to, e.g. b_{1010}.

The first three measures are available as options for itemset evaluation in Christian Borgelt's implementations.

Lift The lift measure was introduced in [19] and compares the conditional probability of an association rule's RHS to its unconditional probability: $lift(X \Rightarrow Y) = \frac{conf(X \Rightarrow Y)}{freq(X)}$. A lift value larger than one denotes evidence of a positive correlation, a value smaller than one negative correlation.

[1] Downloadable at `http://www.borgelt.net/fpm.html`.

Information gain Information gain is best known for choosing the splitting test in inner nodes of decision trees. For a given RHS Y, it measures the reduction of its entropy:

$$H(Y, \mathcal{T}) = -\sum_{b \in \mathcal{T}/\sim_{\{Y\}}} \frac{|b|}{|\mathcal{T}|} \log_2 \frac{|b|}{|\mathcal{T}|}$$

by the presence of the LHS X:

$$IG(X \Rightarrow Y) = H(Y, \mathcal{T}) - \sum_{b \in \mathcal{T}/\sim_{\{X\}}} \frac{|b|}{|\mathcal{T}|} H(Y, b)$$

Normalized χ^2 The χ^2 test is a test for statistical independence of categorical variables. For the two-variable case given by RHS Y and LHS X, occurrence counts can be arranged in a contingency table:

	$Y = 1$	$Y = 0$	
$X = 1$	$\lvert b_{11}\rvert$	$\lvert b_{10}\rvert$	$sup(X)$
$X = 0$	$\lvert b_{01}\rvert$	$\lvert b_{00}\rvert$	$\lvert\mathcal{T}\rvert - sup(X)$
	$sup(Y)$	$\lvert\mathcal{T}\rvert - sup(Y)$	$\lvert\mathcal{T}\rvert$

To derive the χ^2 value, the observed values are compared to the expected value (the normalized product of the margins, e.g. $E_{11} = \frac{sup(X)}{|\mathcal{T}|} \cdot \frac{sup(Y)}{|\mathcal{T}|} \cdot |\mathcal{T}|$):

$$\chi^2(X \Rightarrow Y) = \sum_{i=0}^{1} \sum_{j=0}^{1} \frac{(|b_{ij}| - E_{ij})^2}{E_{ij}}$$

The χ^2 value scales with the number of cases the LHS occurs in. To normalize this value, Borgelt's FIM implementations normalize this value by the support of the LHS.

Multi-way χ^2 Brin *et al.* proposed to use the χ^2 test to evaluate itemsets directly [20]. Each item $i \in I$ is considered its own itemset and instead of a 2×2 contingency table, a multiway table with $2^{|I|}$ cells is populated by the cardinalities of the blocks derived from $\mathcal{T}/\sim_{\{i|i \in I\}}$. The χ^2-value is calculated as in the two-dimensional case. The degrees of freedom for such a table are $df(I) = 2^{|I|} - 1 - |I|$, and if the χ^2 value exceeds a given p-value for that many df, the itemset is considered significant. Brin *et al.* also propose an *interest* measure for individuals cells: $interest(b_v) = |1 - \frac{O_v}{E_v}|$, and propose to consider the combination of item presences and absences of the cell with the highest interest value the most relevant contribution of the found itemset.

Entropy The entropy definition used for a binary variable above can be extended to partitions with more than two blocks and therefore used on the partition induced by the items of an itemset, similar to the preceding treatment of χ^2:

$$H(\{i \mid i \in I\}) = -\sum_{b \in \mathcal{T}/\sim_{\{i|i \in I\}}} \frac{|b|}{|\mathcal{T}|} \log_2 (\frac{|b|}{|\mathcal{T}|})$$

The entropy is highest, equal to $|I|$, if all blocks are equally likely, and 0 if there is only one block. Heikinheimo *et al.* have proposed mining low-entropy itemsets [21].

Maximum Entropy evaluation Not a measure *per se*, De Bie has proposed to use maximum entropy models to sample data sets that conform to certain constraints derived from $\mathcal{T}$, e.g. row and column margins, i.e. support of individual items and sizes of transactions, in the expectation [17]. Found patterns can be reevaluated on these databases and rejected if they occur in more than a certain proportion of them.

3 The Quest generator

The Quest generator was introduced in the paper that jump-started the area of frequent itemset mining (FIM), and arguably the entire pattern mining field [1]. The generative process is governed by a number of parameters:

- L – the number of potentially large itemsets (source itemsets) embedded in the data.
- N – the number of items from which source itemsets can be assembled.
- $|I|$ – the average size of source itemsets.
- $|t|$ – the average size of transactions in the data.
- $|\mathcal{T}|$ – the cardinality of the data set.
- c – the "correlation level" between successive itemsets.

The generator proceeds in two phases: it *first* generates all source itemsets, and in a *second* step assembles the transactions that make up the full data set from these source itemsets. The authors, working in the shopping basket setting, aimed to model the phenomenon that certain items are typically bought together and several such groups of items would make up a transaction. This also means that the output of FIM operations can be compared to the source itemsets to get an impression of how well such mining operations recover the underlying patterns, i.e. the individual typical shopping baskets.

3.1 Source itemset generation

For each of the L source itemsets, the size is sampled from a *Poisson* distribution with mean I. A fraction of the items used in the source itemset formed in iteration i are taken randomly from the itemset formed in iteration $i-1$. This fraction is sampled from an exponential distribution with mean c. The rest of the items are sampled uniformly from N. Each source itemset is assigned a weight, i.e. its probability of occurring in the data, sampled from an exponential distribution with unit mean, and a corruption level, i.e. a probability value for only the partial source itemset embedded into a transaction, sampled from a normal distribution with mean 0.5 and variance 0.1. Source itemsets' weights are normalized so that they sum to 1.0.

3.2 Transaction generation

For each of the D transactions, the size is sampled from a Poisson distribution with mean T. Source itemsets to be embedded into the transaction are chosen according to their weight, and their items embedded according to their corruption level. Importantly, this means that source itemsets are selected independently from each other.

If the number of items to be embedded exceeds the remaining size of the transaction, in half the cases the items are embedded anyway, and the transaction made larger, in the other half of the cases, the transaction is made smaller, and the items transferred for embedding into the succeeding transaction.

4 Related work

The seminal paper on FIM, which also introduced the Quest generator, was published almost twenty years ago [1]. The authors used the generator to systematically explore the effects of data characteristics on their proposed algorithm, using several different transaction and source itemset sizes, evaluating a number of values for the data set cardinality (9 values), the number of items in the data (5 values), and transaction sizes (6 values) while keeping the other parameters fixed, respectively, specifically the number of source itemsets used. It is unclear whether more than data set was mined for each setting, a question that becomes relevant given the probabilistic nature of the correlation, corruption, and source itemset weight effects.

A similar kind of systematic evaluation can still be found in [22], although the authors did not evaluate the effect of different values for N (and also continue to keep L fixed throughout). The evaluation found in [23], however, already limits itself to only two Quest-generated data sets. In line with this trend, the author of [24] used only four Quest-generated data sets which he augmented by three UCI data sets [25], and PUSMB data sets that act as stand-ins for "dense" data sets, i.e. sets with relatively few items coupled with large transaction sizes. The evaluation reported in [26] uses one artificial data set, one UCI data set, and the PUSMB data set.

The systematic use of the Quest generator came to a virtual halt after Zheng *et al.* reported that one of the Quest-generated data sets shows different characteristics from real-life data and that algorithmic improvements reported in the literature did not transfer to real-life data [8]. Notably, the authors pointed out that CLOSET [26] scales worse than CHARM [27], a result that Zaki *et al.* verified in revisiting their work and comparing against CLOSET as well [28], and that runs contrary to the experimental evidence presented in [26] by the authors of CLOSET, probably due to the difference in used data sets.

The typical evaluation of FIM related approaches afterwards consisted of using two Quest-generated data sets, a number of UCI data sets, and the real-life data sets made available to the community, e.g. in the Frequent Itemset Mining Implementation competitions [29, 9]. This has lead to the paradoxical situation

that while techniques for FIM have proliferated, the amount of data sets on which they have been evaluated has shrunk, in addition to a lack of control over these data sets' characteristics. Also, all evaluations limited themselves to evaluating efficiency questions.

In the same period, data sets begun to be characterized by the distribution of the patterns mined from them, starting with [28] and continued in [11, 12, 6, 30, 31]. These analyses have given rise to techniques for "'inverse itemset mining" that, starting from FIM results, generate data leading to the same distribution of mined itemsets. While these data sets could be used for efficiency evaluations, they are dependent on the data from which patterns are mined in the first place, and the lack of clearly defined patterns prevents quality evaluations. In a similar vein falls the generator proposed in [15] which uses the MDL principle to generate data sets that will lead to similar itemsets mined, even though it serves a different purpose, namely to protect the anonymity of original data sources.

Finally, FIM research has spawned a large number of interestingness measures and literature discussing what desirable characteristics of such measures are [32, 33]. It is at present unknown, however, whether any of these measures manages to recover the patterns underlying the data, and the closest research has come to such evaluations are attempts to establish how well interestingness measures for association rules align with domain experts' interest [34, 35]

5 Pattern recovery

The fact that the Quest generator assembles transactions in terms of source itemsets gives us the unique opportunity to compare the output of a frequent itemset mining operation to the original patterns. Note that this is different from the approach taken in [11, 12, 15, 16] – in those works databases were generated that would *result* in the itemsets (or at least of the same number of itemsets of certain sizes) being mined that informed the generating process. Contrary to this, we cannot be sure that the output of the frequent itemset mining operation has any relation to the source patterns from which the data is generated, although we of course expect that that would be the case. To the best of our knowledge, this is the first time that such an objective comparison of mined to source itemsets has been performed.

5.1 Experimental setup

For reasons of computational efficiency, we use only few $(10, 100)$ source itemsets in our experiments. This allows us to mine with relatively high support thresholds without having to expect missing (too many) source itemsets. We generate data with $N = 2000, T = 10, I = 4$, with corruption turned off. We generate 100 data sets for each setting and average the results over them. Since we are not considering run times at this point, we used Apriori in Christian Borgelt's implementation with support threshold $100/L\%$, a generous threshold given that we can expect each transaction to consist on average of $10/4 = 2.5$ itemsets.

This corresponds to a relatively easy setting since the source itemsets have high support and apart from the correlation-induced overlap, items are unlikely to appear in several itemsets.

We mine three types of patterns: frequent itemsets, closed itemsets, and maximal itemsets. While frequent itemsets will be guaranteed to include all source itemsets recoverable at the minimum support threshold, they will also include all of their subsets, and possibly additional combinations of items. Closed itemsets might miss some source itemsets if the probabilistic process of the Quest generator often groups two itemsets together while generating transactions, an effect that should not be very pronounced over 100 data sets, however. On the other hand can closed itemsets be expected to avoid finding subsets of frequent sets unless those are intersections of source itemsets, and to restrict supersets of source itemsets to unions of them. Maximal itemsets, finally, can be expected to consist of unions of source itemsets.

For each pattern type, we filter according to the different measures afterwards:

a) Lift: independence is equivalent to a value of 1, we use a threshold of 1.01.
b) Information gain: independence will manifest as $IG = 0$, therefore we set the threshold to 0.01.
c) Normalized χ^2: the value can lie between 0 and 1, we therefore set the threshold to 0.01.
d) Multi-way χ^2: no threshold is needed for this measure but a significance level, we choose 0.05. A high score does not always indicate that the itemset as a whole is relevant, however. To interpret selected itemsets, we use the block with the highest interest value. To give an example, if $\{i_1, i_2, i_3\}$ attains a high score but the block with highest interest is b_{101}, we interpret $\{i_1, i_3\}$ as the pattern, and i_2 as being negatively correlated with it, hence coming from a different source itemset. For this measure, we can therefore also assess how many negative correlations were identified, and how many of those correctly.
e) Entropy: there is not clear way to set a maximal threshold for entropy. We require for the entropy of itemsets to be at most half of the maximal entropy for a set of that size.
f) Maximum entropy evaluation: we use an empirical p-value of 0.05 to reject the null hypothesis that the items in an itemset are independent from each other, i.e. an itemset must not be frequent on more than 5% of the sampled data sets. We sample 100 times from the maximum entropy model of each data set. We will therefore risk false negatives but evaluating patterns on 10000 data sets already taxes our computational resources.
g) Pairwise χ^2: we also added an additional measure, calculating the χ^2 value for any pair of items in the set, normalizing and requiring a minimum value of 0.01.

5.2 Data sets created without corruption of source itemsets

Pattern counts are shown cumulatively so that the top of the bar corresponds to the total amount of itemsets mined, and vertical show the proportion taken

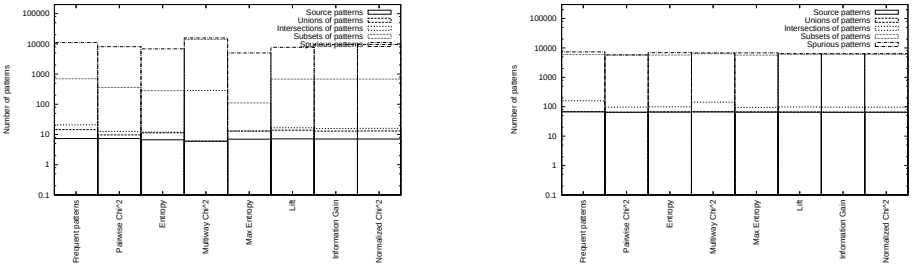

Fig. 1. Results for mining frequent itemsets for $L = 10$ and $L = 100$, no corruption

up by different categories of itemsets. We show the actual itemsets recovered, unions of itemsets, intersections of itemsets, subsets of itemsets that cannot be mapped to intersections, and itemsets that we cannot map to source itemsets at all, labeled "spurious".

Figure 1 shows the results for mining frequent patterns on uncorrupted data, with $L = 10$ on the left-hand side, and $L = 100$ on the right-hand side. The first setting, mining frequent patterns for $L = 10$ without corruption leads to the largest result sets, larger than for $L = 100$, so that filtering for entropy and multi-way χ^2, and evaluation through Maximum Entropy models was not finished by the time of submission and those results are therefore preliminary. As can be seen, most source pattern can be recovered, however a large part of the output is taken up by subsets of the frequent itemsets and since those correspond to correlating items, the additional measures are not able to remove them. It is a bit surprising to see how many spurious itemsets, i.e. combination of items that do not originate from source itemsets, are not removed by the interestingness measures.

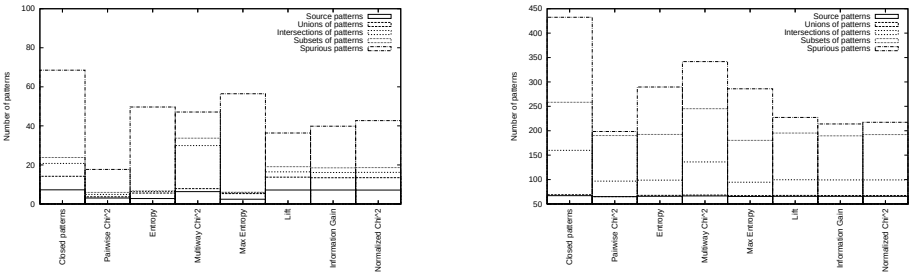

Fig. 2. Results for mining closed itemsets for $L = 10$ and $L = 100$, no corruption

Mining closed patterns removes most of the subsets from the result set and leaves the output much more manageable (Figure 2). We see that pairwise χ^2 is a very aggressive criterion, reducing not only the amount of spurious itemsets

but also rejecting source itemsets, as does the Maximum Entropy evaluation due to the false negative effects. On the other hand are quite a few spurious sets not filtered out by the MaxEnt evaluation. The association rule measures are effective in filtering itemsets, as is multiway χ^2, which has the added advantage that it separates out negative correlations and therefore recovers more intersections of itemsets.

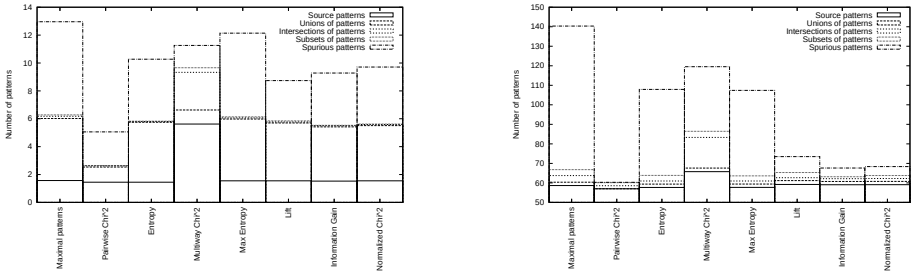

Fig. 3. Results for mining maximal itemsets for $L = 10$ and $L = 100$, no corruption

Mining maximal itemsets, finally, reduces the result set to roughly the amount of source itemsets but only half of those are related to source itemsets, as Figure 3. The trends that can be observed for closed sets hold here as well, with multiway χ^2 recovering additional source itemsets, pairwise χ^2 filtering aggressively, and the other measures reducing non-relevant patterns somewhat.

It is interesting to see that the association rule measures that have not been designed to evaluate itemsets as such are more effective in reducing the result sets than the itemset measures. In the case of multiway χ^2, this seems to be the price for recovering additional patterns by identifying negative correlations.

All measures were used with rather lenient thresholds and stricter thresholds might improve their performance. The problem is, however, that deciding which threshold to use is not straight-forward when working with real life data.

Given that this uncorrupted embedding of source itemsets is the best-case scenario, an obvious next question is what effect pattern corruption has on the output. The authors of [1] motivated such corruption by shoppers that might not need all ingredients of a shopping basket, for instance, but in other settings, e.g. sensorial data, corruption might result from weak signals or malfunctioning equipment.

5.3 Data sets created with corruption of source itemsets

The settings used in the preceding section are relatively easy: source itemsets are embedded without corruption so that they should be well recoverable. In real-life data, however, it is possible that patterns are being corrupted while they are being acquired, making the task of identifying them harder. Since the decision

which items are not embedded is independently made for each item, corruption should occur uniformly and still enable itemset recovery, however.

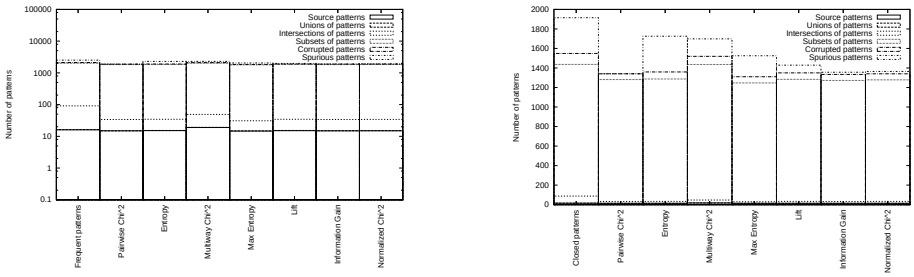

Fig. 4. Results for mining frequent and closed itemsets for $L = 100$, with corruption

As we see in Figures 4, however, the deterioration is significant. Only few itemsets in the result set correspond to source itemsets. Instead, the result set consists mainly of fragments of the source itemsets and while the interestingness measures are effective in filtering out spurious sets, reducing the rest of the result set would be up to the user. Page constraints prevent us from showing the plot for maximal itemsets but the trends we have seen for $L = 10$ and for the other two pattern types for $L = 100$ hold there as well.

As our experiments have shown, FIM results should definitely be taken with a grain of salt. Even by using additional interestingness measures, it is not assured that the mining process recovers the patterns underlying the data.

6 Summary and Conclusions

In this work, we have for the first time evaluated interestingness measures for frequent itemset mining objectively. Due to a lack of data whose underlying patterns are unknown and whose characteristics cannot be easily controlled, it had been currently unknown how effective FIM approaches are in recovering the underlying patterns.

We have revisited to Almaden Quest data generator, and have used the fact that it constructs data from explicit patterns. By generating data sets and performing frequent itemset mining on them, we could compare the mined patterns against the source itemsets used to construct the data. We found not only that mining frequent, closed, or maximal patterns leads to result sets that include many non-relevant patterns in addition to source itemsets but also that several interestingness measures that have been proposed in the literature are only partially effective in reducing the result set to the relevant patterns.

The ramifications of our results could be far-reaching: our experiments call the usefulness of itemset mining results into question since underlying patterns cannot be reliably recovered. Clear-cut evidence for such usefulness could take

the form of implementing gained knowledge in the domains in which the data originate but such evaluations do not exist to the best of our knowledge. Using domain experts to evaluate itemset mining results, while easier and cheaper, has its pitfalls and current interestingness measures might not be up to the task of identifying the relevant itemsets. This is a challenge the community needs to take on to increase the utilization of itemset mining in non-academic contexts.

We have mentioned above that the Quest generator could also be used to generate data sets with different transaction lengths, densities etc. The generator would need to be modified to address the concerns raised by Zheng *et al.* but such a modified generator could then be used to stage more comprehensive run times evaluations as well. We are currently performing such research.

References

1. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules in large databases," in *Proceedings of the 20th VLDB*. Santiago de Chile, Chile: Morgan Kaufmann, Sept. 1994, pp. 487–499.
2. L. Geng and H. J. Hamilton, "Interestingness measures for data mining: A survey," *ACM Comput. Surv.*, vol. 38, no. 3, 2006.
3. N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, "Discovering frequent closed itemsets for association rules." in *ICDT*, LNCS, vol. 1540. Springer, 1999, pp. 398–416.
4. J.-F. Boulicaut and B. Jeudy, "Mining free itemsets under constraints." in *IDEAS '01*, M. E. Adiba, C. Collet, and B. C. Desai, Eds., 2001, pp. 322–329.
5. T. Calders and B. Goethals, "Mining all non-derivable frequent itemsets." in *PKDD*, Springer, 2002, pp. 74–85.
6. K. Gouda and M. J. Zaki, "Genmax: An efficient algorithm for mining maximal frequent itemsets," *Data Min. Knowl. Discov.*, vol. 11, no. 3, pp. 223–242, 2005.
7. Y. Pei and O. Zaane, "A synthetic data generator for clustering and outlier analysis," Tech. Rep., 2006.
8. Z. Zheng, R. Kohavi, and L. Mason, "Real world performance of association rule algorithms," in *KDD*, 2001, pp. 401–406.
9. R. J. Bayardo Jr., B. Goethals, and M. J. Zaki, Eds., *FIMI '04*.
10. B. Liu, W. Hsu, S. Chen, and Y. Ma, "Analyzing the subjective interestingness of association rules," *IEEE Intelligent Systems*, vol. 15, no. 5, pp. 47–55, 2000.
11. G. Ramesh, W. Maniatty, and M. J. Zaki, "Feasible itemset distributions in data mining: theory and application," in *PODS*. ACM, 2003, pp. 284–295.
12. G. Ramesh, M. J. Zaki, and W. Maniatty, "Distribution-based synthetic database generation techniques for itemset mining," in *IDEAS*. IEEE Computer Society, 2005, pp. 307–316.
13. C. Cooper and M. Zito, "Realistic synthetic data for testing association rule mining algorithms for market basket databases," in *PKDD*, LNCS, vol. 4702. Springer, 2007, pp. 398–405.
14. A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas, "Assessing data mining results via swap randomization," *TKDD*, vol. 1, no. 3, 2007.
15. J. Vreeken, M. van Leeuwen, and A. Siebes, "Preserving privacy through data generation," in *ICDM*, IEEE Computer Society, 2007, pp. 685–690.
16. A. Guzzo, D. Saccà, and E. Serra, "An effective approach to inverse frequent set mining," in *ICDM*, IEEE Computer Society, 2009, pp. 806–811.
17. T. D. Bie, "Maximum entropy models and subjective interestingness: an application to tiles in binary databases," *Data Min. Knowl. Discov.*, vol. 23, no. 3, pp. 407–446, 2011.
18. K. McGarry, "A survey of interestingness measures for knowledge discovery," *The Knowledge Engineering Review*, vol. 20, no. 1, pp. 39–61, 2005.
19. S. Brin, R. Motwani, J. D. Ullman, and S. Tsur, "Dynamic itemset counting and implication rules for market basket data," in *SIGMOD Conference*, J. Peckham, Ed. ACM Press, 1997, pp. 255–264.
20. S. Brin, R. Motwani, and C. Silverstein, "Beyond market baskets: Generalizing association rules to correlations." J. Peckham, Ed. ACM Press, 1997, pp. 265–276.
21. H. Heikinheimo, J. K. Seppänen, E. Hinkkanen, H. Mannila, and T. Mielikäinen, "Finding low-entropy sets and trees from binary data," in *KDD*, ACM, 2007, pp. 350–359.
22. M. J. Zaki, "Scalable algorithms for association mining," *IEEE Trans. Knowl. Data Eng.*, vol. 12, no. 3, pp. 372–390, 2000.
23. J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in *SIGMOD Conference*, ACM, 2000, pp. 1–12.
24. M. J. Zaki and C.-J. Hsiao, "ChArm: An efficient algorithm for closed association rule mining," Computer Science Department, Rensselaer Polytechnic Institute, Tech. Rep., October 1999.
25. C. Blake and C. Merz, "UCI repository of machine learning databases," 1998. [Online].
26. J. Pei, J. Han, and R. Mao, "Closet: An efficient algorithm for mining frequent closed itemsets," in *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 2000, pp. 21–30.
27. M. J. Zaki, "Generating non-redundant association rules," in *KDD*, 2000, pp. 34–43.
28. M. J. Zaki and C.-J. Hsiao, "Charm: An efficient algorithm for closed itemset mining," in *SDM*, SIAM, 2002.
29. B. Goethals and M. J. Zaki, Eds., *FIMI '03*, ser. CEUR Workshop Proceedings, vol. 90. CEUR-WS.org, 2003.
30. P. Palmerini, S. Orlando, and R. Perego, "Statistical properties of transactional databases," in *SAC*, ACM, 2004, pp. 515–519.
31. F. Flouvat, F. D. Marchi, and J.-M. Petit, "A new classification of datasets for frequent itemsets," *J. Intell. Inf. Syst.*, vol. 34, no. 1, pp. 1–19, 2010.
32. P.-N. Tan, V. Kumar, and J. Srivastava, "Selecting the right interestingness measure for association patterns," in *KDD*. ACM, 2002, pp. 32–41.
33. T. Wu, Y. Chen, and J. Han, "Re-examination of interestingness measures in pattern mining: a unified framework," *Data Min. Knowl. Discov.*, vol. 21, no. 3, pp. 371–397, 2010.
34. M. Ohsaki, S. Kitaguchi, K. Okamoto, H. Yokoi, and T. Yamaguchi, "Evaluation of rule interestingness measures with a clinical dataset on hepatitis," in *PKDD*, LNCS, vol. 3202. Springer, 2004, pp. 362–373.
35. D. R. Carvalho, A. A. Freitas, and N. F. F. Ebecken, "Evaluating the correlation between objective rule interestingness measures and real human interest," in *PKDD*, Springer, 2005, pp. 453–461.
36. J. Peckham, Ed., *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*. ACM Press, 1997.