SDM14
2014 SIAM International Conference on Data Mining
April 24 – 26, 2014, Philadelphia, Pennsylvania, USA

Reviews For Paper

Paper ID	26
Title	Finding the True Frequent Itemsets

Masked Reviewer ID: Assigned_Reviewer_2

Review:

Question	
Brief summary of the paper's contributions.	In this paper the authors develop a method to compute a new quality threshold tau^ from a user defined frequency threshold tau. The new threshold is used to enumerate an approximate collection of True Frequent Itemsets. This method is based on the assumption that the distribution of frequent itemsets in a real world dataset is different from the distribution of frequent itemsets in the infinitely sized transaction dataset. Moreover the distribution in the infinite dataset is the exact one, and the dataset at hand is a random sample obtained from the prior distribution.
Overall rating	Weak Reject: I vote for rejection, but will not argue if it is accepted.
Brief justification for overall rating.	The paper seems to have good theoretical background, however, the applicability of the method is not convincing enough. Moreover the experimental section should be improved.
Detailed comments. For more information, go to http://www-users.cs.umn.edu/~banerjee/sdm14/guidelines.html.	Typo's ===== 6.2 These datasets differs in size -> differ… 6.5 be found using by our algorithm -> by using… Detailed comments ============ The authors made a nice theoretical study on finding a new frequency threshold to mine an approximate collection of True Frequent Patterns. They use Vapnik-Chervonenkis dimension to obtain a bound on the original threshold such that only True Frequent patterns are obtained. However, is this definition of True Frequency really interesting for a given user? First of all, the number of patterns returned can still be very large (since only the frequency parameter is adapted) and, moreover, many duplicates are still found when using the adapted parameter. Also, FIM is often used as preprocessing technique and the patterns are then used in other frameworks to obtain a vastly reduced collection of patterns. How do the extra false positives/negatives influence these results? Some of the explanations can be hard to follow. Adding pseudo-code could help understand the algorithm for finding the new quality threshold better.

	In the experimental section, the authors explain that 20 random databases are generated per FIMI dataset. They use uniform sampling from the transactions of the datasets, therefore essentially duplicating the smaller datasets. Is this a valid way to test this method? Another approach would be to sample without replacement X datasets of size/X and then using the original dataset as ground truth. Also, how does this method operate on biased databases? I.e., suppose the dataset at hand does not reflect the true generating distribution? In the first experiment, the authors state that for almost every dataset the set of frequent itemsets contains false positives and false negatives. However, each of the runs could include just one false positive/negative, which in the end would not be a bad result. It can therefore be informative to add some numbers on the number of false positive/negatives found by the experiments. It can also be interesting for readers to see the updated frequency threshold to get a feeling of how much the threshold is being adapted.

Masked Reviewer ID: Assigned_Reviewer_3

Review:

Question

Brief summary of the paper's contributions.	This paper proposes a new and interesting concept, True Frequent Itemsets (TFIs), which are extracted from the True Frequent Itemsets (FIs) with at least certain probability in a collection of samples obtained from an unknown probability distribution π on transactions. Furthermore, it also designs an algorithm to identify a threshold such that the collection of itemsets with frequency at least contains only TFIs with probability at least (σ, θ are user-specified). Finally, some experimental results based on some datasets from FIMI'04 repository, were provided to illustrate the better results of the proposed method by contrast with the algorithm using Chernoff and Union bound. Although this paper proposed a new and interesting concept, i.e. True Frequent Itemsets, overall I do not think this concept is a breakthrough in the field of Frequent Itemsets. In fact, there have been existing research works discussing how to obtain effective Frequent Itemsets, which are similar to this paper. Moreover, the experimental results are not enough to show the advantages of this paper when compared to the method of Chernoff and Union bound.
Overall rating	Weak Accept: I vote for acceptance, but leaving it out of the program would be no great loss.

Brief justification for overall rating.	1) This paper presents a new and interesting concept, True Frequent Itemsets, and shows an interesting connection with the Set-Union Knapsack Problem. 2) The paper is well written and is easy to understand. 3) The novelty of the concept, True Frequent Itemsets, is a little limited. Because the True Frequent Itemsets cannot avoid false positives completely, the problem of finding True Frequent Itemsets could be converted to the problem of finding Frequent Itemsets with small fault tolerant, which has been discussed in some recent research works. Furthermore, this paper (in section 4 and section 5) mainly focuses on the usage and applications of previous work, which limits its creativity. 4) In section 6.5, Inclusion of TFIs (Recall) in Experimental evaluation, it is not enough to show the advantage of this paper. It will be more powerful if it provides more contrasted experimental results in the wider range of frequency θ. Although the proposed method shows the better results, it needs to provide more experimental results in "Vanilla" case which will strength the advantage of this paper. 5) A minor syntax error in the experimental result explanation in section 6.3: Table 1 reports, for different datasets the fraction of times that the set contained false positives (FP) and was missing TFIs (false negatives (FN)) over 20 datasets from the same ground truth.
Detailed comments. For more information, go to http://www-users.cs.umn.edu/~banerjee/sdm14/guidelines.html.	1) This paper presents a new and interesting concept, True Frequent Itemsets, and shows an interesting connection with the Set-Union Knapsack Problem. 2) The paper is well written and is easy to understand. 3) The novelty of the concept, True Frequent Itemsets, is a little limited. Because the True Frequent Itemsets cannot avoid false positives completely, the problem of finding True Frequent Itemsets could be converted to the

	problem of finding Frequent Itemsets with small fault tolerant, which has been discussed in some recent research works. Furthermore, this paper (in section 4 and section 5) mainly focuses on the usage and applications of previous work, which limits its creativity. 4) In section 6.5, Inclusion of TFIs (Recall) in Experimental evaluation, it is not enough to show the advantage of this paper. It will be more powerful if it provides more contrasted experimental results in the wider range of frequency θ. Although the proposed method shows the better results, it needs to provide more experimental results in "Vanilla" case which will strength the advantage of this paper. 5) A minor syntax error in the experimental result explanation in section 6.3: Table 1 reports, for different datasets the fraction of times that the set contained false positives (FP) and was missing TFIs (false negatives (FN)) over 20 datasets from the same ground truth.

Masked Reviewer ID: Assigned_Reviewer_4

Review:

Question

Brief summary of the paper's contributions.	Using a probabilistic approach, the authors formulate the notion of true frequent itemsets (TFIs), which are the underlying itemsets that would be generated with high probability, theta. They also describe an algorithm that can be used to pick a threshold, theta, that will eliminate a false positive TFI with probability, 1-delta, where delta is a user specified parameter. The analysis of the algorithm uses an apparently novel approach based on the empirical VC-dimension and the Set-Union Knapsack problem.
Overall rating	Accept: I will argue for acceptance.
Brief justification for overall rating.	The paper seems quite solid from a theoretical point of view and the experimental results lend additional support to the validity of the theoretical work. In addition, the analysis of the proposed algorithm for finding TFIs seems quite novel, at least in the association analysis domain.
Detailed comments. For more information, go to http://www-users.cs.umn.edu/~banerjee	This paper defines the problem of mining True Frequent Itemsets, (TFIs), which are the underlying itemsets that would be generated with high probability, theta. They also describe an algorithm that can be used to pick a threshold, theta, that will eliminate a false positive TFI with probability, 1-delta, where delta is a user specified parameter. The analysis of the algorithm seems very unique for association

/sdm14 /guidelines.html.	analysis, being carried out in terms of the empirical VC-dimension and the Set-Union Knapsack problem. A bound is derived that can be computed by the used of linear programming. Experiments were conducted to test the algorithm. The algorithm retrieves a high percentage of TFIs with no false positives. The proposed method is compared to another one, also proposed by the authors. The algorithm is only described in text. It would be nice to see a more implementable description or a pointer to actual code so that others can further explore this approach.

SDM14
2014 SIAM International Conference on Data Mining
April 24 – 26, 2014, Philadelphia, Pennsylvania, USA

Meta-Reviews For Paper

Paper ID	26
Title	Finding the True Frequent Itemsets

Masked Meta-Reviewer ID: Meta_Reviewer_2

Meta-Reviews:

Question	
Overall rating	Weak Accept: I vote for acceptance, but leaving it out of the program would be no great loss.
Justification of the overall rating	The paper presents a PAC learning like framework for evaluating item set mining algorithms, based on a probalistic framework. By appealing to the notion of VC-dimension and the computational problem of Set-Union Kapsack, the paper proposes an approach to finding an approximately true frequent itemsets in that framework. This is a long awaited direction of research to place some statistical rigor into the area of frequent itemset mining.