On the Estimation of Frequent Itemsets for Data Streams: Theory and Experiments

Pierre-Alain Laur[1] Richard Nock[1] Jean-Emile Symphor[1]
[1]U. Antilles-Guyane, Martinique, France
{palaur,rnock,je.symphor}@martinique.univ-ag.fr

Pascal Poncelet[2]
[2]LG2IP-Mines d'Alès, France
pascal.poncelet@ema.fr

ABSTRACT

In this paper, we devise a method for the estimation of the true support of itemsets on data streams, with the objective to maximize one chosen criterion among {precision, recall} while ensuring a degradation as reduced as possible for the other criterion. We discuss the strengths, weaknesses and range of applicability of this method that relies on conventional uniform convergence results, yet guarantees statistical optimality from different standpoints.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval; I.5.3 [Pattern Recognition]: Clustering

General Terms: Algorithms.

Keywords: Data stream mining.

1. INTRODUCTION

A growing body of works arising from researchers in Data bases and Data Mining deals with data arriving in the form of continuous potentially infinite streams, *i.e.* an ordered sequence of item occurrences that arrives in timely manner. Data streams have seen the emergence of crucial problems for databases that were previously not as pregnant, such as the accurate retrieval of informations in a data flow that prevents its exact storage, and whose information may evolve through time. We consider *items* to be the unit information, and *itemsets* to be sets of items. An itemset is θ-*frequent* (frequent for short) if it occurs in at least a fraction θ of the data stream, called its support. An important task is to build the set of the most frequent items or itemsets encountered. The subject of this paper is to propose a tight and extensive study of the application of local uniform convergence results [4] to extend classical supports to *statistical* supports.

2. THEORETICAL STATEMENT

There are ***two*** sources of error for the estimation of frequent itemsets: it is possible that some itemsets observed as frequent might in fact not be frequent anymore from a longer observation of the data stream, even without a drift of the observation odds; on the other hand, some itemsets observed as not frequent may well in fact be frequent from a longer history of the data stream. The point is that it is statistically hard to nullify both sources of error from the observation of a *subset*, even very large, of the whole data stream [4]. This unsatisfiable goal can be relaxed to the tight control of one source or error, while keeping the other one within reasonable bounds: the user chooses a source of error, and fixes some related parameters; then, the source of error chosen is nullified with high probability, while the other one incurs a loss as limited as possible. We now formalize this.

The data stream is supposed to be obtained from the repetitive sampling of a potentially huge *domain* X which contains all possible itemsets. Each itemset is sampled independently through a distribution $\mathcal{D}$, on which we make *no* distribution assumption, except that it remains fixed through time. Our problem, as explained in [1], is to recover $X^*_\theta = X_\theta \cap S^*$, where $0 \leq \theta \leq 1$. Here, $X_\theta = \{T \in X : \rho_X(T) \geq \theta\}$, with $\rho_X(T) = \sum_{T' \in X : T \leq_t T'} \mathcal{D}(T')$, and $T \leq_t T'$ means that itemset T is a subset of T'. Furthermore, S^* is the set of itemsets observed from the stream (S), augmented by conventional generalization algorithms [2]. Now, $\forall T \in S^*$, we cannot compute exactly $\rho_X(T)$, since we do not know X and $\mathcal{D}$. Rather, we have access to its best unbiased estimator $\rho_S(T)$, which can be easily computed from S: $\forall T \in S^*, \rho_S(T) = \sum_{T' \in S : T \leq_t T'} w(T')$, with $w(T')$ the weight (observed frequency) of T' in S. A simple and computationally attractive way to address our problem is to solve the following problem: find some $0 < \theta' < 1$ and approximate as best the set X^*_θ by the set of *observed* θ'-frequent of S^*, that is: $S^*_{\theta'} = \{T \in S^* : \rho_S(T) \geq \theta'\}$. This finally amounts to fixing an accurate value for θ'. Statistically speaking, it is hard to find some θ' that nullifies the overall error, $X^*_\theta \Delta S^*_{\theta'}$ for any m. However, it is possible to obtain some fairly strong constraints on its two components, $|X^*_\theta \backslash S^*_{\theta'}|$ and $|S^*_{\theta'} \backslash X^*_\theta|$, *i.e.* the basis of the two sources of errors outlined in the beginning of this Section, and these constraints hold *regardless* of m. To model them, we use the conventional definitions of precision and recall [1]; Maximizing the precision `P` is equivalent to minimizing our first source of error. Symmetrically, maximizing the recall `R` is equivalent to minimizing our second source of error.

We adopt the concise probabilistic notation of [3], and define for some predicate P the notation $\forall^\delta P$ which means that P holds for all but a fraction $\leq \delta$ of the sets S sampled under distribution $\mathcal{D}$. The following definition is the cornerstone of our approach.

Definition 1. $\forall 0 \leq \theta \leq 1$, $\forall 0 \leq \varepsilon \leq 1$, $\forall S \subseteq X$, we say that S^* is a $\mathbf{sup}{-}(\theta,\varepsilon){-}\mathbf{cover}$ of X iff $\forall T \in X^*_\theta$, $\rho_S(T) \geq \rho_X(T) - \varepsilon$. Respectively, we say that S^* is an $\mathbf{inf}{-}(\theta,\varepsilon){-}\mathbf{cover}$ of X iff $\forall T \in S^* \backslash X^*_\theta$, $\rho_S(T) \leq \rho_X(T) + \varepsilon$.

CIKM'05, October 31–November 5, 2005, Bremen, Germany.
ACM 1-59593-140-6/05/0010.

The way we use definition 1 is simple. Consider that the user has fixed both the theoretical support $0 \leq \theta \leq 1$, and a *statistical risk* parameter $0 < \delta < 1$. Suppose we can find ε such that: $\forall^{\delta}, S^*$ is an inf-(θ, ε)-cover of X. Now, fix $\theta' = \theta + \varepsilon$, so that we keep $S^*_{\theta+\varepsilon}$. We observe $\forall T \in S^* \backslash X^*_{\theta}, \rho_S(T) \leq \rho_X(T) + \varepsilon < \theta + \varepsilon$. Thus, we obtain $\forall^{\delta}, S^*_{\theta+\varepsilon} \subseteq X^*_{\theta}$, which easily yields: $\forall^{\delta}, \mathtt{P} = 1$. Thus, there is *no* first source of error, with high probability. Symmetrically, suppose we can find ε such that $\forall^{\delta}, S^*$ is a sup$-(\theta, \varepsilon)-$cover of X, and fix this time $\theta' = \theta - \varepsilon$, so that we keep $S^*_{\theta-\varepsilon}$. Because of the property of S^*, we observe $\forall T \in X^*_{\theta}, \rho_S(T) \geq \rho_X(T) - \varepsilon \geq \theta - \varepsilon$, which yields $\forall^{\delta}, X^*_{\theta} \subseteq S^*_{\theta-\varepsilon}$, and finally: $\forall^{\delta}, \mathtt{R} = 1$, *i.e.* there is *no* second source of error with high probability. Our problem is thus reduced to finding an accurate value of ε such that S^* is a sup or inf $-(\theta, \varepsilon)-$cover of X with high probability. The following Theorem gives a value ε which yields with high probability a sup$-(\theta, \varepsilon)-$cover of X.

THEOREM 1. *$\forall X, \forall \mathcal{D}, \forall m > 0, \forall 0 \leq \theta \leq 1, \forall 0 < \delta \leq 1$, the following holds: $\forall^{\delta}, S^*$ is a sup$-(\theta, \varepsilon)-$cover of X, for any ε satisfying: $\varepsilon \geq \sqrt{(1/(2m)) \ln(|X^*_{\theta}|/\delta)}$. Respectively, $\forall^{\delta}, S^*$ is an inf$-(\theta, \varepsilon)-$cover of X, for any ε satisfying: $\varepsilon \geq \sqrt{(1/(2m)) \ln(|S^* \backslash X^*_{\theta}|/\delta)}$.*

Theorem 1 says that finding (inf/sup)$-(\theta, \varepsilon)-$covers is a fairly easy task $\forall m$. The following argument shows that there are no significant better covers. Informally, we build a skewed distribution $\mathcal{D}$ on some very simple X^*_{θ}, such that with probability $\geq \delta$ we "miss" the (θ, ε)-cover for some value of ε slightly smaller than those of Theorem 1.

THEOREM 2. *$\exists X, \exists \mathcal{D}, \exists m > 0, \exists 0 \leq \theta \leq 1, \exists 0 < \delta \leq 1$ such that the following holds: with probability $\geq \delta$, S^* is **not** a sup$-(\theta, \varepsilon)-$cover of X, for any ε satisfying: $\varepsilon \leq c\sqrt{(1/(2m)) \ln(|X^*_{\theta}|/\delta)}$. Respectively, with probability $\geq \delta$, S^* is **not** an inf$-(\theta, \varepsilon)-$cover of X, for any ε satisfying: $\varepsilon \leq c\sqrt{(1/(2m)) \ln(|S^* \backslash X^*_{\theta}|/\delta)}$. Here, c is some constant < 1.*

Theorem 2 says that the criterion which is not controlled incurs a loss which is, in one sense, also statistically near-optimal; a simple argument shows that the *value* of this loss behaves in a very reasonable manner.

We now shift to a discussion on the way our approach behaves when there is a *distribution drift*, *i.e.* when $\mathcal{D}$ changes through time. It turns out that our approach can be tailored in a very simple way to estimating these changes in X^*_{θ}. This simply consist in estimating $\rho_S(.)$ on the basis of a *moving window*, wide enough to ensure m large enough, and regularly sampling the data stream. All other parameters *do not change*. Figure 1 explains that, with this straightforward adaptation, the distribution drift is estimated with respect to the moving average of the distributions (thick lines), and *not* with respect to the true distributions (regular line). We estimate for any itemset T the fluctuations of a moving average $\overline{\rho}_X(T)$ instead of $\rho_X(T)$. With respect to this change, it is straightforward to show that the results still hold under *any* distribution drift, to keep maximal precision or recall with respect to the *average* drift. This smoothes the small local drifts, but keeps the significant variations of $\mathcal{D}$ within the detection range.

There only remains to upperbound $|X^*_{\theta}|$ and $|S^* \backslash X^*_{\theta}|$ to compute empirically ε for Theorem 1. Since $|X^*_{\theta}| + |S^* \backslash X^*_{\theta}| = m^*$, we shall use afterwards in the experiments the same upperbound, m^*, for both cardinals.

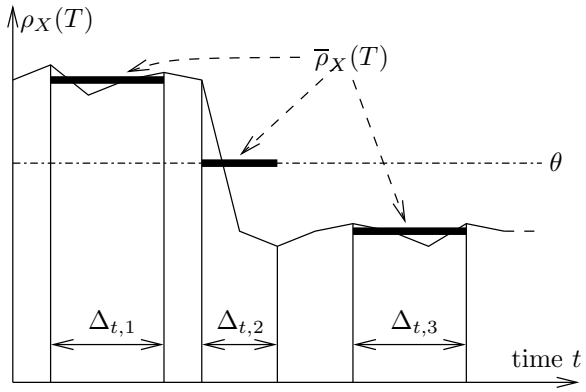

Figure 1: A moving window makes it possible to track distribution drifts. In this example, we may detect that T is θ-frequent during window $\Delta_{t,1}$ while it is not θ-frequent anymore during $\Delta_{t,3}$.

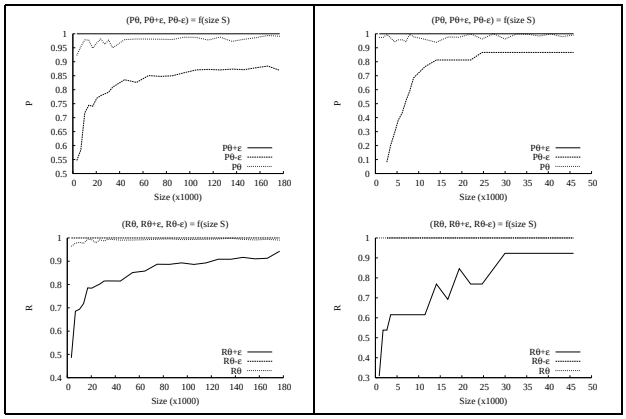

Figure 2: Plots for 2 of our databases, Accidents (left, $\theta = .6$) and Retail (right, $\theta = .07$), with $\delta = .05$. Precision (top) and recall (bottom) are given for the three methods, picking $S^*_{\theta-\varepsilon}, S^*_{\theta}, S^*_{\theta+\varepsilon}$. x-axis=m.

3. EXPERIMENTS

Experiments are reported in Figure 2, where each point is an average over 10 runs of the setup of [1], with the same data bases. This clearly confirms the theory, as the parameter controlled is always at its maximum, *regardless* of m, while the conventional method of keeping S^*_{θ} virtually always fails to this goal. This tends to confirm the robustness of the method, as it also holds for a broad range of supports on each of the data bases we have used.

4. REFERENCES

[1] P.-A. Laur, J.-E. Symphor, R. Nock, and P. Poncelet. Statistical Supports for Frequent Itemsets on Data Streams. In P. Perner and I. Atsushi, editors, *Machine Learning and Data Mining in Pattern Recognition*. Springer Verlag LNCS 3587 (to appear), 2005.

[2] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowlede discovery. *Data Mining and Knowledge Discovery*, 1:241–258, 1997.

[3] D. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37:355–363, 1999.

[4] V. Vapnik. *Statistical Learning Theory*. John Wiley, 1998.