Average Number of Frequent (Closed) Patterns in Bernouilli and Markovian Databases

Loïck LHOTE, François RIOULT, Arnaud SOULET
GREYC, CNRS - UMR 6072, Université de Caen Basse-Normandie
F-14032 Caen cedex, France
{forename.surname}@info.unicaen.fr

Abstract

In data mining, enumerate the frequent or the closed patterns is often the first difficult task leading to the association rules discovery. The number of these patterns represents a great interest. The lower bound is known to be constant whereas the upper bound is exponential, but both situations correspond to pathological cases. For the first time, we give an average analysis of the number of frequent or closed patterns. Average analysis is often closer to real situations and gives more information about the role of the parameters. In this paper, two probabilistic models are studied: a BERNOULLI *and a* MARKOVIAN. *In both models and for large databases, we prove that the number of frequent patterns,* for a fixed frequency threshold *, is exponential in the number of items and polynomial in the number of transactions. On the other hand,* for a proportional frequency threshold *, the number of frequent patterns is* polynomial in the number of items *and does not involve the number of transactions. Finally, we prove in the* BERNOULLI *model that the number of closed patterns, for a proportional frequency threshold, is polynomial in the number of items.*

1 Introduction

The aim of the data mining is to extract relevant information from large databases. These databases gather *transactions*, which are sets of *items*, for example a list of purchase in a shop or a set of biomedical characteristics. In association rules discovery, the sets of items which are frequently present play a central role. These frequent *patterns* can tell that a conjunction of items are often present together in the transactions. They allow to build association rules, which are at the core of numerous data processes, but the difficulty of the task is known to lie in the frequent patterns mining.

The motivation of our article is to make clearer this difficulty. When the complexity of a pattern mining algorithm is studied, it most of the time focuses on the worst cases, or on the bottleneck of specific sub-tasks, such as database accesses. For example, in A-PRIORI [2], the first and most popular algorithm, there will be as much database scans as the size of the largest frequent pattern. Facing to large databases, this is an important point. More precisely, GUNOPULOS et al. have shown that mining the frequent patterns with such level-wise algorithms requires as many database accesses as there are elements in the positive and negative borders [5] (i.e. the number of maximal frequent and minimal infrequent patterns). It follows that the complexity of A-PRIORI intuitively lies in the quantity of frequent patterns. The number of database accesses does not address the real difficulty of pattern mining, and TOIVONEN gives in [11] a probabilistic method based on sampling, which finds all the probably frequent patterns with only one database access.

In order to improve the efficiency of pattern mining, *closed* patterns are very interesting. A closed pattern is the maximal pattern (w.r.t. the inclusion) of the set of patterns having the same frequency and being present in the same transactions. When they are associated with the corresponding pattern of transactions, both constitute a *concept*. Conceptual learning is a hot topic [12], and closed patterns are an easy way to non redundant association rules [13]. Their mining has then been widely examined [7, 3].

Gunopulos et al. have shown that deciding whether there exists a frequent pattern with t items is NP-complete [6, 9]. The associate counting problem is #P-hard. When all the items belong to all the transactions, each pattern is frequent pattern so that, in the worst case, the number of frequent patterns is of order $O(2^m)$. On the other hand, if each transaction is empty, all the patterns are infrequent and the number of frequent patterns is of constant order. There is a large gap between the worst and the best cases and both situations are pathological examples. In practice, the real order is not known.

Regarding the complexity, average analysis is an interesting point of view for three reasons. First of all, each

database is associated to a probability, so that the analysis considers the diversity of cases. Then, if the model is close to the reality, this analysis gives a realistic average behaviour of the studied parameter, which is sometimes far from the worst case. Finally, if the concentration property around the average is satisfied, fast counting algorithm are available. On the other hand, real life is often complex and the modelization is not an easy task. Furthermore, the more complex is the modelization, the more difficult is the average analysis.

We found one such study [9], but it is related to the failure rate of A-PRIORI. It is useful for predicting the number of candidates that the algorithm will have to check. This work confirms the results of [4], which used an upper bound. On other hand, the authors of the A-PRIORI algorithm have explained in [1] that there are very few long patterns in a random database. In a previous work [8], we used the same probabilistic model and recall here our results. We also improve them with studying a MARKOVIAN model.

In this article, we propose for the first time, an average analysis of the number of frequent and closed patterns for two probabilistic models of databases. In the worst cases, the results are well mastered: there is an exponential number of frequent patterns, according to the number of items. On the contrary, we provide an *average analysis* of the number of frequent patterns, with two probabilistic models: the first is the simple BERNOULLI, where each transaction contains an item with a uniform probability p; the second is a MARKOVIAN model, which better reflects the correlation of the real data, because each transaction is generated by a MARKOVIAN source. Indeed, the model considers local correlations between close items but the transactions stay independent two by two. Both models are rather simple and not close to the reality, however they point very interesting phenomena. In particular, we show that the number of frequent patterns, with a proportional minimum frequency threshold, is surprisingly polynomial in the number of items and does not involve the number of transactions.

2 Database modelization

2.1 Definitions and notations

The set of items is noted $\mathcal{I} = \{1, \ldots, m\}$ and has the cardinality m whereas the set of transactions is noted $\mathcal{T} = \{t_1, \ldots, t_n\}$ and has the cardinality n. By definition, each transaction t_i is a subset of $\mathcal{I}$ and a database $\mathcal{B}$ is a couple $(\mathcal{I}, \mathcal{T})$. The database $\mathcal{B}$ admits a matrix representation χ where χ is a boolean matrix whose coefficients $\chi_{i,j}$, $i = 1..n$, $j = 1..m$ satisfy $\chi_{i,j} = 1$ if and only if $j \in t_i$ We will not discuss here about the methods for obtaining such a boolean matrix, starting from continuous or multi-valued items (see [10] for an example). The *support* of a pattern $A \subset \mathcal{I}$ is the set of all transactions containing A, and the frequency of A is the size of its support.

Definition 1 (frequent pattern). *Let $\mathcal{B}$ be a binary database with m items and n transactions and γ a strictly positive integer smaller than n. A pattern A is said γ-frequent if $|support(A)| \geq \gamma$.*

We now give, in this framework, the definition of a γ-closed pattern:

Definition 2 (frequent closed pattern). *Let $\mathcal{B} = (\mathcal{I}, \mathcal{T})$ be a binary database with m items and n transactions and γ a strictly positive integer smaller than n. Fix also χ the matricial representation of $\mathcal{B}$. A pattern A is γ-closed if:*
- A is a γ-frequent pattern,
- for all item j in $\mathcal{I} \backslash A$, there exists a transaction t_i in $support(A)$ such that j does not belong to t_i, i.e., $\chi_{i,j} = 0$.

2.2 First hypothesis

Hypothesis on the sizes: biological databases have many items and few transactions, leading to *fat* databases, which are wider than high. It was observed that the number of frequent or closed patterns have a completely different behaviours than in *large* databases, which are higher than wide. It is then normal to consider this phenomenon. In this article, we deal with *large databases* where the number of items is at most polynomial in the number of transactions and vice versa. The mathematical version of this property is: $(\mathcal{H}_1) \qquad \log m \sim c \log n, \quad c > 0.$
Recall that m (resp. n) is the number of items (resp. transactions). We will see that this property plays an important role in our results.

Hypothesis on the frequency threshold: a pattern is said frequent as soon as its frequency rises over a user defined threshold γ. For $\gamma = 1$, experiments lead to an exponential number of frequent/closed patterns whereas for $\gamma = n$, it is constant and often equal to zero. Hence according to the frequency threshold, we may expect different results. Two kinds of thresholds are considered in this article:
- fixed threshold: γ is fixed and does not depend on n or m. It corresponds in practice to very small threshold compared to the number n of transactions (10 for instance).
- proportional threshold: there exists $r \in]0, 1[$ such that $\gamma = r \cdot n$. In this case, the threshold is a non-negligible proportion of all the transactions (10% for instance).

Remark that with the proportional threshold, the number of unfrequent patterns is much more important than with the fixed threshold, leading to a complete different behavior (see next section).

2.3 BERNOULLI model

We now describe the simple BERNOULLI model. Since we can not appreciate *in advance* the correlations existing in a databases, we first suppose that:

Model 1 (BERNOULLI model). *The database $(\chi_{i,j})_{i=1..n,j=1..m}$ forms an independent family of random variables which follows the same* BERNOULLI *law of parameter p in $]0,1[$.*

This model is far from the reality. Indeed, an equivalent in Information Theory is to modelize the French language with a memoryless source that respects the probability of each letter. The result is not very realistic but theoretical analysis can yet be lead.

2.4 MARKOVIAN model

In the second model, each transaction is a sequence of m random variables, with values in $\{0,1\}$ that follow a MARKOVIAN process of order k. In other words, an item belongs to a transaction according to a law that only involves the values of the k previous items.

The MARKOVIAN model (of order k) is completely described by the way the first k items are affected and the transition probabilities $(p_{w\to x})_{w,x}$, $x \in \{0,1\}, w \in \{0,1\}^k$, where $p_{w\to x}$ is the probability that the new item take the value x knowing that the k previous items form the word w. We suppose that the initial values of the first k variables are given by the distribution $f_{init} = (f_w)_{w\in\{0,1\}^k}$.
We now precisely describe the second model.

Model 2 (MARKOVIAN model). *Fix $k \geq 1$, $f_{init} = (f_w)_{w\in\{0,1\}^k}$ an initial distribution on $\{0,1\}^k$ and $(p_{w\to x})_{w\in\{0,1\}^k, x\in\{0,1\}}$ the transition probabilities. Then, each transaction is computed independently from the other transactions according to the following method:*
for a transaction $t = (\chi_1, \ldots, \chi_m)$, $(\chi_1, \ldots, \chi_k)$ is computed according to the initial distribution f_{init}. Then, the values of $\chi_{k+1}, \ldots, \chi_m$ are sequentially evaluated using the k previous values and the transition probabilities.

Contrary to the BERNOULLI model, the MARKOVIAN model introduces local correlations between items. This is of course insufficient for modeling real databases but it constitutes an improvement compare to the first model. MARKOVIAN databases may have a sense if an organisation of the items entails that close items are much more correlated than distant items. In bioinformatics, the items are the genes and it is known that close genes are much more correlated than distant genes.

3 Theoretical results

This section enumerates three new theorems about the average number of frequent or closed patterns in a random database. Results in the BERNOULLI model always involve explicit constants. On the other hand, results with the MARKOVIAN model express with theoretical constants related to the transition matrix.

Theorem 1 (Fixed threshold). *Fix a threshold γ and suppose that the hypothesis $\mathcal{H}_1$ is fulfilled. Then, the average number of frequent patterns $F_{m,n,\gamma}$ in a* BERNOULLI *database of parameter p or in a* MARKOVIAN *database with matrix transition P is polynomial in the number of transactions and exponential in the number of items,*

$$F_{m,n,\gamma} \sim c_0 \frac{n^\gamma}{\gamma!}\theta^m, \qquad \theta > 1.$$

In the BERNOULLI *model, $\theta = 1 + p^\gamma$ and $c_0 = 1$. In the* MARKOVIAN *model, θ is the dominant eigenvalue of a strictly positive matrix and c_0 is related to dominant spectral objects.*

This average result is standard with the intuition given by the worst case: according to the number of items, there is an exponential quantity of frequent patterns. This phenomena is well known, and the difficulty of pattern mining lies in the size of the item set $\mathcal{I}$. This theorem also shows that the number of frequent patterns polynomially depends on the number of transactions. This is coherent with the complexity of A-PRIORI regarding the number of database scans, equal to the maximum pattern length. The results with a proportional frequency threshold are more surprising:

Theorem 2 (Proportional threshold). *Fix $r \in]0,1[$ and suppose that the hypothesis $\mathcal{H}_1$ is fulfilled. Finally, suppose that the frequency threshold satisfies $\gamma = r \cdot n$. Then, the average number of frequent patterns $F_{m,n,\gamma}$ in a* BERNOULLI *database of parameter p or in a* MARKOVIAN *database with matrix transition P is at most polynomial in the number in the number of items with an upper bound that does not involve the number of transactions,*

$$F_{m,n,\gamma} \leq c_1 m^s.$$

In the BERNOULLI *model, it is an equivalence as soon as r is not a power of p. Then $c_1 = 1/s!$ and $s = \lfloor \log r / \log p \rfloor$.*

Theorem 2 is an unexpected result. Indeed, experiments usually highlight a very important number of frequent patterns, even with such a proportional threshold. It is nevertheless not sufficient to conclude that this quantity is exponential. When the threshold γ varies from 1 to n, the number of frequent patterns goes from an exponential behaviour

to a constant one. Theorem 2 suggests that a proportional threshold is sufficient to get a polynomial behaviour.

We finish this section with a theorem concerning the closed patterns in the BERNOULLI model.

Theorem 3 (Closed patterns). *For* $\gamma > (1+\epsilon)\dfrac{\log m}{|\log p|}$*, the number of closed patterns,* $C_{m,n,\gamma}$ *and frequent patterns are equivalent,*

$$C_{m,n,\gamma} \sim F_{m,n,\gamma}.$$

With the uncorrelated model of BERNOULLI, the number frequent and closed patterns coincide. In real databases, closed patterns are yet well known to be very few than frequent patterns and allow to design very efficient mining algorithms because they synthesise the correlations in the data. Mining the closed pattern is more complicated than for the frequent patterns, and it is not justified for the uncorrelated databases.

Corollary 1. *With a proportional threshold* $\gamma = r \cdot n$*, the average number of closed patterns* $C_{m,n,\gamma}$ *is polynomial in the number of items:*

$$C_{m,n,\gamma} \sim \frac{m^s}{s!} \text{ where } s = \lfloor \log r / \log p \rfloor.$$

For a large enough frequency threshold γ, the number of frequent patterns is almost equal to the number of closed patterns. Experiments with classical synthetic databases `T10I4D100K` and `T40I10D100K` confirm the theoretical result [8] since synthetic data are almost without correlations. On the other hand, the theorem does not reflect the real behaviour of `Pumsb` (another dtabase concerning fault diagnosis problem of electro-mechanical devices), because this dataset does not follow our uncorrelated model and the same remark applies with the base `Connect`.

4 Conclusion

In this article, we have given three new results about the average number of frequent and closed patterns in random databases. It is the first time that such analysis is performed. Two probabilistic models were studied: a BERNOULLI model that generates uncorrelated databases and a MARKOVIAN model, where close items are correlated. These models are far from real life but give new fruitful information for the pattern mining. In particular, we proved that for a fixed frequency threshold, the number of frequent patterns is polynomial in the number of transactions and exponential in the number of items. On the other hand, if the frequency threshold is proportional to the number of transactions, the average number of frequent patterns admits a polynomial behaviour. This last result is unexpected for specialists that commonly refer to the worst case and its exponential growth.

The average number of frequent patterns is quite interesting to evaluate the complexity of A-PRIORI. In order to be complete, we also need to evaluate the size of the negative border, and it is a work in progress. In the same field, we are also interested in many other open problems, such as the number of closed patterns for a fixed threshold and for other probabilistic models, the average size of the positive border, the average size of the largest frequent (which corresponds to the number of steps for A-PRIORI), etc.

References

[1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo. Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*, pages 307–328, 1996.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In *International Conference on Very Large Data Bases (VLDB'94), Santiago de Chile*, pages 487–499, 1994.

[3] H. Fu and E. Mephu Nguifo. How well go lattice algorithms on currently used machine learning testbeds ? In *First International Conference on Formal Concept Analysis*, 2003.

[4] F. Geerts, B. Goethals, and J. Van den Bussche. A tight upper bound on the number of candidate patterns. In *ICDM*, pages 155–162, 2001.

[5] D. Gunopulos, H. Mannila, R. Khardon, and H. Toivonen. Data mining, hypergraph transversals, and machine learning. In *PODS 1997*, pages 209–216, 1997.

[6] D. Gunopulos, H. Mannila, and S. Saluja. Discovering all most specific sentences by randomized algorithms. In *ICDT*, pages 215–229, 1997.

[7] S. O. Kuznetsov and S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal of Experimental and Theoretical Artificial Intelligence*, 14(2-3):189–216, 2002.

[8] L. Lhote, F. Rioult, and A. Soulet. Average number of frequent and closed patterns in random databases. In *Conférence d'Apprentissage, CAp'05*, pages 345–360, 2005.

[9] P. W. Purdom, D. V. Gucht, and D. P. Groth. Average-case performance of the apriori algorithm. *SIAM Journal on Computing*, 33(5):1223–1260, 2004.

[10] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data*, pages 1–12. ACM Press, 1996.

[11] H. Toivonen. Sampling large databases for association rules. In *International Conference on Very Large Data Bases*, pages 134–145. Morgan Kaufman, 1996.

[12] R. Wille. Concept lattices and conceptual knowledge systems. In *Computer mathematic applied, 23(6-9):493-515*, 1992.

[13] M. J. Zaki. Generating non-redundant association rules. In *SIGKDD'00, Boston*, pages 34–43, 2000.