

Algorithm for discovery of sequential patterns in big data and application on price of brent crude oil

T. Ruzgas¹, J. Arnastauskaite^{1,2} and I. Juskaite¹

¹Department of Applied Mathematics, Kaunas University of Technology,
Kaunas, Lithuania:
Email : tomas.ruzgas@ktu.lt, indre.juskaite@ktu.edu

²Department of Computer Sciences, Kaunas University of Technology,
Kaunas, Lithuania:
Email: jurgita.arnastauskaite@ktu.lt

ABSTRACT

Frequent sequence mining in large population is important for economics, engineering, genetics and many other databases analysis. Exact algorithms are designed to search for frequent sequence, they repeatedly re-selects the entire population. If the population is large, the search is slow or requires very large computer resources. The paper proposes a new probabilistic frequencies sequence finding algorithm which analyze a randomized sample of the primary population concluded in a specific way. Based on this analysis, statistical inferences are made about frequent sequence in the primary population. This algorithm is not accurate, but it runs much faster than the exact algorithms and is suitable for exploratory statistical analysis. Probabilities of errors are estimated by statistical methods. The probabilistic algorithm can be combined with the exact algorithms of frequent sequences search.

Keywords: Generalized sequential pattern algorithm, probabilistic algorithm, probability of errors.

Mathematics Subject Classification: 62L10, 60A99.

Computing Classification System: I.4

1. INTRODUCTION

As machine learning progresses, many exact algorithms have been proposed for the analysis of frequent sequence. The most popular exact algorithm is GSP (Srikant and Agrawal, 1995a; Srikant and Agrawal, 1995b), other popular exact algorithms are SPADE (Zaki, 2001), SPAM (Ayres et al., 2002), PrefixSpan (Pei et al., 2001), FreeSpan (Han et al., 2000). If the population is large, searching for frequent sequences using precision algorithms is slow or requires large computer resources. In some problems, the ascertainment of frequent sequences with certain estimated error is acceptable, which makes it possible to apply less accurate probabilistic algorithms. Probabilistic algorithms are much faster than exact algorithms because, instead of performing multiple scans of the primary population, they analyse a smaller sample of data generated in certain way. Based on this analysis, statistical inferences are made about frequent sequences in the primary population. ProMFS (Hryniow and Dzielinski, 2012) is one of the approximate probabilistic frequent sequence finding algorithms.

This algorithm generates a new shorter sequence based on the statistical characteristics of the main sequence. Another approximate algorithm is ApproxMAP (Kum et al., 2003). The basic idea behind the ApproxMAP algorithm is to find the sequence used in many other sequences instead of the exact sequence search. The iBiRM (Seki and Yoon, 2019) algorithm includes a new computational method adapted for obtaining recurrent sequences in order to reduce the number of population scans.

The probabilistic algorithm for finding frequent sequences proposed in the paper analyses specific random sample of the primary population. Based on this analysis, statistical inferences are made about frequent sequences in the primary population. The probability of errors are estimated by statistical methods.

2. ALGORITHMS

2.1. Generalized sequential pattern algorithm

Generalized sequential pattern (GSP) algorithm is an exact algorithm for detecting frequent sequence at a selected frequency threshold $\epsilon \in (0, 1)$. Suppose the original population is a numbered data set $S = (S_1, S_2, \dots, S_q)$. Set S elements S_i may gain m different values $x_1, x_2, \dots, x_m$. Sequence $x_{i_1}, x_{i_2}, \dots, x_{i_k}$ is called frequent, if

$$P(x_{i_1}, x_{i_2}, \dots, x_{i_k}) = \frac{1}{q} \sum_{j=1}^q 1_{[S_j = x_{i_1}, S_{j+1} = x_{i_2}, \dots, S_{j+k-1} = x_{i_k}]} \geq \epsilon$$

Alternatively, this sequence is called rare.

GSP algorithm selects the entire primary population, determines which sequences are rare and doesn't investigate them further. During the first re-election of data GSP algorithm reads the first level (one character) sequences $x_1, x_2, \dots, x_m$ and determines which sequences are frequent. The second level (two character) candidates are then formed from the determined frequency sequences $x_1x_1, x_1x_2, \dots, x_1x_m, x_2x_1, \dots, x_2x_m, \dots, x_mx_1, \dots, x_mx_m$, which may be common. Potential candidates for frequent sequences do not include sequences that have already been identified as rare sequences. Evidently, if a sequence is rare, then all sequences that have that rare sequence will also be rare, for instance, if x_1x_2 is rare, then sequence $x_1x_2x_1, x_1x_2x_2$ and so forth as well are rare. Similarly, common sequences are set for other levels. The algorithm ends when there are no more candidates in ordinary level for frequent sequences. The GSP algorithm determines exactly which sequences are frequent at the selected threshold ϵ , however, it makes multiple primary sample scans. If the primary sample is large, then the calculation time of the GSP algorithm is high because many population scans are needed.

2.2. Probabilistic algorithm

The probabilistic algorithm is much faster than the GSP because it doesn't analyse the entire primary data sequence, but rather its shorter random sample. The probabilistic algorithm is approximate, but its probability of error can be estimated.

The random sample U of primary sequence is constructed as follows:

- Generate a random variable Y , which gains values $1, 2, \dots, q$ with equal probabilities $1/q$, sequence of realizations $Y_1, Y_2, \dots, Y_n$.
- When looking for first level (one character) frequent sequences, a random sample of U elements x_i are simply $U_{Y_1}, U_{Y_2}, \dots, U_{Y_n}$. The second level of random sampling of element pairs $x_i x_j$ is $(U_{Y_1}, U_{Y_1+1}), (U_{Y_2}, U_{Y_2+1}), \dots, (U_{Y_n}, U_{Y_n+1})$. k -level random sample for elements sets $x_i x_j$ is $(U_{Y_1}, \dots, U_{Y_1+k-1}), (U_{Y_2}, \dots, U_{Y_2+k-1}), (U_{Y_n}, \dots, U_{Y_n+k-1})$ and so forth. Such a sample is made up of a return sample because some of the numbers Y_i may be repeated. A non-return random sample is formed from the recurring numbers Y_i by eliminating all recurring numbers and additionally generating new numbers until a set of non-recurring numbers is obtained $Y_1, Y_2, \dots, Y_n$.

Using the GSP algorithm, we determine the sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ empirical frequencies in a random sample U :

$$\phi(x_{t_1}, x_{t_2}, \dots, x_{t_n}) = \frac{1}{n} \sum_{j=1}^n 1_{\{U_{Y_j} = x_{t_1}, U_{Y_j+1} = x_{t_2}, \dots, U_{Y_j+k-1} = x_{t_k}\}}$$

Choose a number $\delta \geq 0$ ($0 \leq \phi - \delta \leq \phi + \delta \leq 1$). Classify sequences $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ in to three classes:

- if $\phi(x_{t_1}, x_{t_2}, \dots, x_{t_n}) \leq \phi - \delta$, then sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ is classified as rare sequence class;
- if $\phi(x_{t_1}, x_{t_2}, \dots, x_{t_n}) \geq \phi + \delta$, then sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ is classified as frequent sequence class;
- if $\phi - \delta < \phi(x_{t_1}, x_{t_2}, \dots, x_{t_n}) < \phi + \delta$, then sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ is classified as intermediate sequence class;

We will discuss error probability estimation for the probabilistic algorithm. Let's fix any sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$. There are two types of errors: I) a sequence that is assigned to a class of frequent sequences, however actually it is rare; II) a sequence that is assigned to a class of rare sequences, however actually it is frequent.

Let's mark $p = \phi(x_{t_1}, x_{t_2}, \dots, x_{t_n})$, $\bar{p} = 1 - p$. Then the probability of the first type of error does not exceed $P(\bar{p} - p > \delta)$, and the probability of the second type of error does not exceed $P(p - p < -\delta)$. It is convenient to use such scheme to estimate these probabilities. Let's define random variables $Z_i = 1$, if $U_{Y_i} = x_{t_1}, U_{Y_i+1} = x_{t_2}, \dots, U_{Y_i+k-1} = x_{t_k}$, $i = 1, \dots, n$, otherwise $Z_i = 0$.

Because of the way of sequencing $Y_1, Y_2, \dots, Y_n$, the random variables $Z_1, Z_2, \dots, Z_n$ are independent and equally distributed, the mean $E Z_i = p$, and dispersion $D Z_i = p(1 - p)$. Probabilities of error can be estimated by standard mathematical statistics methods: based on the properties of the binomial distribution for the refundable sample and the properties of the hypergeometric distribution for the non-refundable sample.

Let's define a random variable $\theta = Z_1 + \dots + Z_n$. If the sample is non-refundable, then the distribution function of the statistics θ is

$$F(l, M) = \sum_{i=0}^l \frac{C_M^l C_{q-M}^{n-l}}{C_q^n}, l = 0, 1, \dots, M$$

where $\sum_{j=1}^q \mathbf{1}_{[x_{t_1}^{(j)}, x_{t_1}^{(j)} + 1 - x_{t_2}^{(j)}, \dots, x_{t_n}^{(j)} + 1 - x_{t_n}^{(j)}]}$. Let's fix $\alpha \in (0, 1)$ and define the numbers A and B as follows: A is the smallest integer satisfying the inequality $F(\theta - 1, A) \geq 1 - \alpha$; B is the largest integer satisfying the inequality $F(\theta - 1, B) \leq \alpha$.

The integers A and B are the lower and upper $(1 - \alpha)$ confidence limits corresponding to the parameter M . Thus, the following inequalities are satisfied:

$$P(M \geq A) = P\left(p(x_{t_1}, x_{t_2}, \dots, x_{t_n}) \geq \frac{A}{q}\right) \geq 1 - \alpha$$

$$P(M \leq B) = P\left(p(x_{t_1}, x_{t_2}, \dots, x_{t_n}) \leq \frac{B}{q}\right) \geq 1 - \alpha$$

If the sample size n is large enough, asymptotic estimators of the probabilities of error can be used, based on a normal approximation of the normalized sums of independent random variables. Based on the central limit theorem with all $v \in (-\infty, \infty)$

$$P\left(\frac{\theta - \mathbb{E}\theta}{\sqrt{D\theta}} < v\right) = \Phi(v) + \Delta_n(v)$$

where Φ is a function of the standard normal distribution and

$$\Delta_n \equiv \sup_v |\Delta_n(v)| \rightarrow 0, n \rightarrow \infty,$$

After marking $w_n(p) = \frac{\sqrt{n}}{\sqrt{p(1-p)}}$, we get representation for error probability estimator

$$\begin{aligned} P(\theta - p > \delta) &= 1 - P(\theta - p \leq \delta) = 1 - P(w_n(\theta - p) \leq w_n \delta) = 1 - P\left(\frac{\theta - \mathbb{E}\theta}{\sqrt{D\theta}} \leq w_n(p) \delta\right) \\ &= 1 - \Phi(w_n(p) \delta) + \Delta_n(w_n(p) \delta) \leq 1 - \Phi(w_n(\sigma_0) \delta) \end{aligned}$$

where $\sigma_0 = \min\left(\sigma, \frac{1}{2}\right)$, and

$$P(\theta - p < -\delta) = P(w_n(\theta - p) < -w_n(p) \delta) = \Phi(-w_n(p) \delta) + \Delta_n(-w_n(p) \delta) \leq \Phi(-w_n(\sigma_1) \delta)$$

where $\sigma_1 = \max\left(\sigma, \frac{1}{2}\right)$.

The Berry–Esseen inequality can be used to estimate the error of the normal approximation

$$\Delta_n \leq \frac{C \mathbb{E}|Z_1 - \mathbb{E}Z_1|^3}{\sqrt{n} (DZ_1)^{3/2}} = \frac{C (1-p)^2 + p^2}{\sqrt{n} \sqrt{p(1-p)^3}}$$

here the universal constant $C = 0.7655$. The actual frequency p is usually unknown. Thus, applying this inequality requires an additional evaluation of the lower boundary $p(1-p)$.

If $\sigma - \delta < p < \sigma + \delta$, then the decision of assigning sequence is not accepted, since the likelihood of assigning sequence error may be high. The probability of error depends on how much the true frequency p differs from ε . Suppose that $p = \varepsilon$. Based on the central limit theorem, $P(\theta \geq \sigma) \rightarrow \frac{1}{2}$ and

$P(\bar{p} < \frac{1}{2}) \rightarrow \frac{1}{2}$, when $n \rightarrow \infty$. Therefore, only by re-electing the entire primary population will we be able to determine whether the sequence $x_{t_1}, x_{t_2}, \dots, x_{t_n}$ is frequent or rare.

On the other hand, whatever $p \in (0, 1)$ is, it is close to the empirical probability $\bar{p}$, when n is large enough because once again based on the central limit theorem with all $\delta > 0$: $P(|\bar{p} - p| > \delta) \rightarrow 0$, $n \rightarrow \infty$. The probability of an event $\bar{p} - \delta < \bar{p} < \bar{p} + \delta$ can be reduced by reducing δ , but then the probabilities of the first and second type of error increases. They can be reduced by increasing n . Thus, the compatibility between δ and n is necessary, and their connection can be expressed with equality $\delta \sqrt{n} = \text{const.}$

3. EMPIRICAL EXAMPLE

Brent crude oil prices USD/bbl. hourly data from January 3, 2010 to January 31, 2020 (data taken from Bloomberg). Number of population elements $q = 88392$, and its elements may acquire these different meanings {A, B, C}: A – if the hourly end price i is higher than the hourly start price; B – if the hourly end price of i is lower than the hourly start price; C – if the hourly end price i is equal to the hourly start price.

To analyse the primary data sequence S , we will apply GSP and probabilistic algorithms and compare the real frequencies determined by the GSP algorithm with the empirical frequencies determined by the probabilistic algorithm. A sequence is considered to be frequent if its true frequency is not lower than 0.2, i.e. $\varepsilon=0.2$.

We choose the size of the random sample 100 and 2000 along with $\delta=0,02$.

Evaluating the likelihood of the first type error, we obtain:

$$Err^{(1)} = 1 - \Phi\left(\frac{\sqrt{n}}{\sqrt{p(1-p)}} \delta\right)$$

where $Err^{(1)} = 1 - \Phi(0.5) \approx 0.3085$, when $n = 100$ and $Err^{(1)} \approx 0.0127$, when $n = 2000$.

When evaluating the likelihood of a second type error, we obtain:

$$Err^{(2)} = \Phi(-2\delta\sqrt{n})$$

where $Err^{(2)} = \Phi(-0.4) \approx 0.3446$, when $n = 100$ and $Err^{(2)} \approx 0,0368$, when $n = 2000$.

Normal approximation error Δ_n , based on Berry-Esseen inequality and assumption $p(1-p) \geq 0,05$, satisfies inequality

$$\Delta_n \leq \frac{C}{\sqrt{n}} \frac{1}{\sqrt{p(1-p)}} \leq \frac{3.424}{\sqrt{n}},$$

Thus $\Delta_n \leq 0.3424$ when $n = 100$ and $\Delta_n \leq 0.0766$ when $n = 2000$.

The results of the empirical study are presented in table 1.

1 table. Results of GSP and probabilistic algorithms

level	subsequence	GSP		$n = 100$		$n = 2000$	
		true probability	class	empirical probability	class	empirical probability	class
1	A	0.548	frequent	0.48	frequent	0.53	frequent
1	B	0.451	frequent	0.51	frequent	0.46	frequent
1	C	0.001	rare	0.01	rare	<0.01	rare
2	AA	0.253	frequent	0.23	frequent	0.24	frequent
2	AB	0.258	frequent	0.21	intermediate	0.27	frequent
2	BA	0.259	frequent	0.33	frequent	0.26	frequent
2	BB	0.204	intermediate	0.17	rare	0.21	intermediate
3	AAA	0.115	rare	0.11	rare	0.10	rare
3	AAB	0.121	rare	0.10	rare	0.12	rare
3	ABA	0.128	rare	0.08	rare	0.14	rare
3	ABB	0.114	rare	0.11	rare	0.11	rare
3	BAA	0.121	rare	0.15	rare	0.12	rare
3	BAB	0.118	rare	0.15	rare	0.11	rare
3	BBA	0.113	rare	0.11	rare	0.12	rare
3	BBB	0.086	rare	0.08	rare	0.09	rare

4. CONCLUSION

The paper presents a probabilistic algorithm for finding frequent sequences in a given sample that analyses a random sample of the primary population and draws statistical inferences about the frequent sequences in the primary population. The probabilistic algorithm is not accurate, but it runs much faster than the exact algorithms and is suitable for exploratory statistical analysis. The probabilities of errors of probabilistic algorithm are estimated by statistical methods. When increasing the random sample, it reduces the probability of error.

5. REFERENCES

- Ayres J., Flannick J., Gehrke J. and Yiu T. 2002. *Sequential pattern mining using a bitmap representation*. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 429–435.
- Han J., Pei J., Mortazavi-asl B., Chen Q., Dayal U. and Hsu M.Ch. 2000. *Freespan: Frequent pattern-projected sequential pattern mining*. In Proc. Knowledge Discovery and Data Mining, pp. 355–359.
- Hryniow K. and Dzielinski A. 2012. *Probabilistic Sequence Mining : Evaluation and Extension of ProMFS Algorithm for Real-Time Problems*, International Journal of Electronics and Telecommunications, 58 (4), pp. 323–326.
- Kum H.Ch., Pei J., Wang W. and Duncan D. 2003. *Approxmap: Approximate mining of consensus sequential patterns*. In Proceedings of the 2003 SIAM International Conference on Data Mining (SIAM DM '03), pp. 311–315.
- Pei J., Han J., Mortazavi-asl B., Pinto H., Chen Q., Dayal U. and M.Ch. Hsu M.-Ch. 2001. *Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth*. In Proc. 17th International Conference on Data Engineering ICDE2001, pp. 215–224.

Srikant R. and Agrawal R. 1995a. *Mining sequential patterns*. In *Proceedings ICDE'95*. Taipei (Taiwan).

Seki H. and Yoon S. 2019. *A New Algorithm for Mining Recurrent Rules from a Sequence Database*, 2019 IEEE International Conference on Systems, Man and Cybernetics, Bari, Italy, pp. 1932-1937.

Srikant R. and Agrawal R. (1995b). *Mining Sequential Patterns: Generalizations and Performance Improvements*. IBM Almaden Research Center.

Zaki M.J. 2001. Spade: *An efficient algorithm for mining frequent sequences*. Machine Learning Journal, 42(1–2):31–60.