Lietuvos matematikos rinkinys. LMD darbai
51 tomas, 2010, 313–318

ISSN 0132-2818
www.mii.lt/LMR/

Tikimybinis dažnų posekių paieškos algoritmas

Julija Pragarauskaitė, Gintautas Dzemyda

Matematikos ir informatikos institutas
Akademijos g. 4, LT-08663 Vilnius

E. paštas: julija.pragarauskaite@gmail.com; dzemyda@ktl.mii.lt

Santrauka. Dažnų posekių paieška didelėse duomenų bazėse yra svarbi biologinių, klimato, finansinių ir daugelio kitų duomenų bazių analizei. Tikslieji algoritmai, skirti dažnų posekių paieškai, daug kartų perrenka visą duomenų bazę. Jeigu duomenų bazė didelė, tai paieška yra lėta arba reikalingi superkompiuteriai. Straipsnyje pasiūlytas naujas tikimybinis dažnų posekių paieškos algoritmas, kuris analizuoja tam tikru būdu sudarytą pradinės duomenų bazės atsitiktinę imtį. Remiantis šia analize daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje. Šis algoritmas nėra tikslus, tačiau veikia daug greičiau negu tikslieji algoritmai ir tinka žvalgomajai statistinei analizei. Klaidų tikimybės įvertinamos statistiniais metodais. Tikimybinis algoritmas gali būti derinamas su tiksliaisiais dažnų posekių paieškos algoritmais.

Raktiniai žodžiai: dažnų sekų paieška, tikimybinis algoritmas, duomenų gavyba.

Įvadas

Pastaraisiais metais pasiūlyta daug tiksliųjų algoritmų dažnų posekių analizei. Populiariausias tikslusis algoritmas GSP [8, 9], kiti populiarūs tikslieji algoritmai yra SPADE [11], SPAM [1], PrefixSpan [7], FreeSpan [3]. Jeigu duomenų bazė didelė, tai dažnų posekių paieška naudojant tiksliuosius algoritmus, yra lėta arba reikalingi superkompiuteriai. Kai kuriuose uždaviniuose dažnų posekių nustatymas su tam tikra įvertinta paklaida yra priimtinas, todėl galima taikyti tikimybinius algoritmus. Tikimybiniai algoritmai yra daug greitesni nei tikslieji, nes, užuot atlikę daugybinius pradinės duomenų bazės nuskaitymus, jie analizuoja tam tikru būdu generuotą trumpesnę duomenų imtį. Remiantis šia analize, daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje.

ProMFS [10] yra vienas iš apytikslių tikimybinių dažnų posekių paieškos algoritmų. Šis algoritmas generuoja naują trumpesnę seką, remdamasis statistinėmis pagrindinės sekos charakteristikomis. Kitas apytikslis algoritmas yra ApproxMAP [5]. Pagrindinė ApproxMAP algoritmo idėja yra vietoje tikslios posekių paieškos rasti posekius, apytiksliai naudojamus daugelyje kitų posekių.

Straipsnyje pasiūlytas tikimybinis dažnų posekių paieškos algoritmas analizuoja tam tikrą pradinės duomenų bazės atsitiktinę imtį. Remiantis šia analize daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje. Klaidų tikimybės įvertinamos statistiniais metodais.

GSP agoritmas

GSP (*Generalized Sequential Pattern mining algorithm*) yra tikslus algoritmas dažniems posekiams nustatyti pagal pasirinktą dažnio slenkstį $\varepsilon \in (0,1)$. Tarkime, kad pradinė duomenų bazė yra sunumeruota duomenų aibė $S = (S_1, S_2, \ldots, S_n)$. Aibės S elementai S_i gali įgyti m skirtingų reikšmių $a_1, a_2, \ldots, a_m$. Posekis $a_{i_1}, a_{i_2}, \ldots, a_{i_k}$ vadinamas dažnu, jeigu

$$p(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) = \frac{1}{N}\#\{j\colon S_j = a_{i_1}, S_{j+1} = a_{i_2}, \ldots, S_{j+k-1} = a_{i_k}\} \geqslant \varepsilon.$$

Antraip, šis posekis yra vadinamas retu.

GSP algoritmas perrenka visą pradinę duomenų bazę, nustato, kurie posekiai yra reti ir jų toliau netiria. GSP algoritmas pirmojo duomenų perrinkimo metu nuskaito pirmojo lygio (vieno simbolio) posekius $a_1, a_2, \ldots, a_m$ ir nustato, kurie posekiai yra dažni. Toliau iš nustatytų dažnų posekių formuojami antrojo lygio (dviejų simbolių) kandidatai $a_1a_2, a_1a_2, \ldots, a_1a_n, a_2a_1, \ldots, a_2a_n, a_na_1, \ldots, a_na_n$, kurie gali būti dažni. Į potencialius dažnų posekių kandidatus nepatenka posekiai, sudaryti iš jau nustatytų retų posekių. Akivaizdu, kad jeigu posekis retas, tai visi posekiai, turintys šį posekį, taip pat bus reti, pavyzdžiui, jei a_1a_2 yra retas, tuomet posekiai $a_1a_2a_1, a_2a_1a_2$ ir t.t. taip pat yra reti. Panašiai nustatomi ir kitų lygių dažni posekiai. Algoritmas baigiamas, kai eiliniame lygyje nebėra kandidatų į dažnus posekius. GSP algoritmas tiksliai nustato, kurios sekos yra dažnos pagal pasirinktą slenkstį ε, tačiau daro daugybinius pradinės duomenų bazės nuskaitymus. Jeigu pradinė duomenų bazė yra didelė, tuomet GSP algoritmo laiko sąnaudos yra didelės, nes tenka daug kartų nuskaityti duomenų bazę.

Tikimybinis algoritmas

Tikimybinis algoritmas yra daug spartesnis nei GSP, nes analizuoja ne visą pradinę duomenų seką, o daug trumpesnę jos atsitiktinę imtį. Tikimybinis algoritmas yra apytikslis, tačiau jo paklaidų tikimybes galima įvertinti.

Pradinės sekos atsitiktinė imtis $\bar{S}$ sudaroma taip:

- Generuojame atsitiktinio dydžio η, įgyjančio reikšmes $1, 2, \ldots, N$ su vienodomis tikimybėmis $\frac{1}{N}$, realizacijų seką $\eta_1, \eta_2, \ldots, \eta_n$.
- Ieškant pirmojo lygio (vieno elemento) dažnų posekių, atsitiktinė imtis $\bar{S}$ elementams a_i yra tiesiog $S_{\eta_1}, S_{\eta_2}, \ldots, S_{\eta_n}$. Antrojo lygio atsitiktinė imtis elementų poroms a_ia_j yra $(S_{\eta_1}, S_{\eta_1+1})$, $(S_{\eta_2}, S_{\eta_2+1})$, $\ldots$, $(S_{\eta_n}, S_{\eta_n+1})$. k-ojo lygio atsitiktinė imtis elementų rinkiniams $a_i \ldots a_k$ yra $(S_{\eta_1}, \ldots, S_{\eta_1+k-1})$, $(S_{\eta_2}, \ldots, S_{\eta_2+k-1})$, $\ldots$, $(S_{\eta_n}, \ldots, S_{\eta_n+k-1})$ ir t. t. Tokia imtis yra sudaryta grąžintiniu ėmimu, nes kai kurie skaičiai η_i gali pasikartoti. Negrąžintiniu ėmimu sudaryta atsitiktinė imtis formuojama iš pasikartojančių skaičių η_i pašalinant visus pasikartojančius skaičius, bei papildomai generuojant naujus skaičius, kol bus gautas nesikartojančių skaičių rinkinys $\eta_1, \eta_2, \ldots, \eta_n$.

Pasinaudoję GSP algoritmu, nustatome posekių $a_{i_1}, a_{i_2}, \ldots, a_{i_k}$ empirinius dažnius atsitiktinėje imtyje $\bar{S}$:

$$\bar{p}_n(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) = \frac{\#\{j\colon S_{\eta_j} = a_{i_1}, S_{\eta_j+1} = a_{i_2}, \ldots, S_{\eta_j+k-1} = a_{i_k}\}}{n}.$$

Pasirenkame skaičių $\delta > 0$ $(0 < \varepsilon - \delta < \varepsilon + \delta < 1)$, $k = 1, 2, \ldots$. Posekius $a_{i_1}, a_{i_2}, \ldots, a_{i_k}$ klasifikuojame į tris grupes: 1) jeigu $\bar{p}_n(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \geqslant \varepsilon + \delta$, tai posekį $a_{i_1}, \ldots, a_{i_k}$ priskiriame dažnų posekių klasei; 2) jeigu $\bar{p}_n(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \leqslant \varepsilon - \delta$, tai posekį $a_{i_1}, \ldots, a_{i_k}$ priskiriame retų posekių klasei; 3) jeigu $\bar{p}_n(a_{i_1}, a_{i_2}, \ldots, a_{i_k}) \in (\varepsilon - \delta, \varepsilon + \delta)$, tai posekį $a_{i_1}, \ldots, a_{i_k}$ priskiriame tarpinių posekių klasei.

Aptarsime tikimybinio algoritmo klaidų tikimybių įvertinius. Fiksuokime kokį nors posekį $a_{i_1}, \ldots, a_{i_k}$. Galimos dviejų rūšių klaidos: 1) posekis priskirtas dažnų posekių klasei, tačiau iš tikro jis yra retas; 2) posekis priskirtas retų posekių klasei, tačiau iš tikro jis yra dažnas.

Pažymėkime $\bar{p}_n = \bar{p}_n(a_{i_1}, \ldots, a_{i_k})$, $p = p(a_{i_1}, \ldots, a_{i_k})$. Tada pirmosios rūšies klaidos tikimybė neviršija $P(\bar{p}_n - p > \delta)$, o antrosios rūšies klaidos tikimybė neviršija $P(\bar{p}_n - p < -\delta)$. Vertinant šias tikimybes, patogu pasinaudoti tokia schema. Apibrėžkime atsitiktinius dydžius $Z_i = 1$, jeigu $S_{\eta_i} = a_{i_1}, S_{\eta_i+1} = a_{i_2}, \ldots, S_{\eta_i+k-1} = a_{i_k}$, $i = 1, \ldots, n$, $Z_i = 0$, priešingu atveju.

Dėl sekos $\eta_1, \eta_2, \ldots, \eta_n$ sudarymo būdo atsitiktiniai dydžiai $Z_1, Z_2, \ldots, Z_n$ yra tarpusavyje nepriklausomi ir vienodai pasiskirstę, vidurkis $\mathbf{E}Z_i = p$, o dispersija $\mathbf{D}Z_i = p(1-p)$.

Klaidų tikimybes galima įvertinti standartiniais matematinės statistikos metodais: remiantis binominio skirstinio savybėmis grąžintinės imties atveju bei hipergeometrinio skirstinio savybėmis negrąžintinės imties atveju [2, pp. 220–224].

Apibrėžkime atsitiktinį dydį

$$\Sigma_n + Z_1 + Z_2 + \cdots + Z_n.$$

Jei imtis negrąžintinė, tai statistikos Σ_n pasiskirstymo funkcija

$$F(l, M) = \sum_{i=0}^{l} \frac{C_M^i C_{N-M}^{n-i}}{C_N^n}, \quad l = 0, 1, \ldots, M,$$

čia $M = \#\{j\colon S_j = a_{i_1}, S_{j+1} = a_{i_2}, \ldots, S_{j+k-1} = a_{i_k}\}$. Fiksuokime $a \in (0, 1)$ ir apibrėžkime skaičius $\underline{M}$ ir $\overline{M}$ taip: $\underline{M}$ yra mažiausias sveikasis skaičius, tenkinantis nelygybę $F(\Sigma_n - 1, \underline{M}) \geqslant 1 - a$; $\overline{M}$ yra didžiausias sveikasis skaičius, tenkinantis nelygybę $F(\Sigma_n - 1, \overline{M}) \leqslant a$.

Sveikieji skaičiai $\underline{M}$ ir $\overline{M}$ yra parametro M apatinis ir viršutinis $(1-a)$ pasikliauties rėžiai. Taigi tenkinamos šios nelygybės:

$$P\{M \geqslant \underline{M}\} = P\left\{p(a_{i_1}, \ldots, a_{i_k}) \geqslant \frac{\underline{M}}{N}\right\} \geqslant 1 - a,$$

$$P\{M \leqslant \overline{M}\} = P\left\{p(a_{i_1}, \ldots, a_{i_k}) \leqslant \frac{\overline{M}}{N}\right\} \geqslant 1 - a.$$

Jei imties didumas n yra pakankamai didelis, galima pasinaudoti asimptotiniais klaidų tikimybių įvertiniais, grindžiamais nepriklausomų atsitiktinių dydžių normuotų sumų normaliąja aproksimacija. Remiantis centrine ribine teorema [4, pp. 264–271], su visais $x \in (-\infty, +\infty)$

$$P\left(\frac{\Sigma_n - \mathbf{E}\Sigma_n}{\sqrt{\mathbf{D}\Sigma_n}} \leqslant x\right) = \Phi(x) + \Delta_n(x),$$

čia Φ yra standartinio normaliojo skirstinio $N(0,1)$ pasiskirstymo funkcija ir

$$\Delta_n \equiv \sup_x \left|\Delta_n(x)\right| \to 0, \quad n \to \infty.$$

Pažymėję $\tau_n = \frac{\sqrt{n}}{\sqrt{p(1-p)}}$, gauname klaidų tikimybių įvertinių išraiškas

$$P(\bar{p}_n - p > \delta) = 1 - P(\bar{p}_n - p \leqslant \delta) = 1 - P\big(\tau_n(\bar{p}_n - p) \leqslant \delta\tau_n\big) =$$
$$= 1 - P\left(\frac{\Sigma_n - \mathbf{E}\Sigma_n}{\sqrt{\mathbf{D}\Sigma_n}} \leqslant \delta\tau_n\right) = 1 - \Phi(\delta\tau_n) + \Delta_n(\delta\tau_n)$$

ir

$$P(\bar{p}_n - p < -\delta) = P\big(\tau_n(\bar{p}_n - p) < -\delta\tau_n\big) = \Phi(-\delta\tau_n) + \Delta_n(-\delta\tau_n).$$

Normaliosios aproksimacijos paklaidos įvertinimui galima taikyti, pavyzdžiui, Berry–Esseeno nelygybę

$$\Delta_n \leqslant \frac{c_0}{\sqrt{n}} \frac{\mathbf{E}|Z_1 - \mathbf{E}Z_1|^3}{(\mathbf{D}Z_1)^{3/2}} = \frac{c_0}{\sqrt{n}} \frac{(1-p)^2 + p^2}{\sqrt{p(1-p)}},$$

čia universalioji konstanta $c_0 = 0{,}7655$. Priminsime, kad mūsų uždavinyje tikrasis dažnis p yra nežinomas. Taigi, taikant šią nelygybę, dar reikia $p(1-p)$ apatinio įverčio.

Jeigu $\bar{p}_n \in (\varepsilon - \delta, \varepsilon + \delta)$, tai prieskyros sprendimas nepriimamas, nes prieskyros klaidos tikimybė gali būti didelė. Prieskyros klaidos tikimybė priklauso nuo to, kiek skiriasi tikrasis dažnis p nuo ε. Tarkime, kad $p = \varepsilon$. Remiantis centrine ribine teorema, $P(\bar{p}_n \geqslant \varepsilon) \to \frac{1}{2}$ ir $P(\bar{p}_n < \varepsilon) \to \frac{1}{2}$, kai $n \to \infty$. Taigi tik perrinkę visą pradinę duomenų bazę galėsime nustatyti, ar posekis $a_{i_1}, \ldots, a_{i_k}$ yra dažnas ar retas.

Kita vertus, kad ir koks būtų $p \in (0,1)$, jis yra artimas empiriniam dažniui $\bar{p}_n$, kai n yra pakankamai didelis, nes vėl remiantis centrine ribine teorema su visais $\mu > 0$: $P(|\bar{p}_n - p| > \mu) \to 0, \quad n \to \infty$. Įvykio $\bar{p}_n \in (\varepsilon - \delta, \varepsilon + \delta)$ tikimybę galima sumažinti mažinant δ, tačiau tada didėja pirmosios ir antrosios prieskyros kaidų tikimybės. Jas galima sumažinti didinant n. Taigi būtinas δ ir n suderintumas, o jų sąryšį galima išreikšti lygybe $\delta\sqrt{n} = \text{const}$.

Eksperimentas

Tirsime finansinių duomenų bazę – valiutų EUR-USD poros valandinius duomenis nuo 2000 m. sausio 3 iki 2010 m. gegužės 7 dienos (duomenys paimti iš Online Trading Platform MetaTrader 4 History Center). Finansinės duomenų bazės elementų skaičius $N = 64074$, o jos elementai gali įgyti šias skirtingas reikšmes $\{A, B, C\}$: A – jeigu i-osios valandos pabaigos kursas yra didesnis nei valandos pradžios kursas; B – jeigu i-osios valandos pabaigos kursas yra mažesnis nei valandos pradžios kursas; C – jeigu i-osios valandos pabaigos kursas yra lygus valandos pradžios kursui.

Pradinei duomenų sekai S tirti taikysime GSP ir tikimybinį algoritmus bei palyginsime tikruosius dažnius, nustatytus GSP algoritmu, su empiriniais dažniais, nustatytais tikimybiniu algoritmu. Laikysime, kad posekis yra dažnas, jeigu jo tikrasis dažnis ne mažesnis nei 0,2, t. y. $\varepsilon = 0{,}2$.

Pasirenkame atsitiktinės imties didumą $n = 100$, $n = 500$ ir $n = 2000$ bei $\delta = 0{,}2$.

Vertindami pirmosios rūšies klaidos tikimybę, gauname

$$1-\Phi(\delta\tau_n)\leqslant 1-\Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right)\equiv\gamma_n^{(1)},$$

čia $\gamma_n^{(1)}\approx 1-\Phi(0{,}5)\approx 0{,}3085$, kai $n=100$, $\gamma_n^{(1)}\approx 1-\Phi(1{,}118)\approx 0{,}1318$, kai $n=500$, ir $\gamma_n^{(1)}\approx 1-\Phi(1{,}118)\approx 0{,}0127$, kai $n=2000$.

Vertindami pirmosios antrosios rūšies klaidos tikimybę, gauname

$$\Phi(-\delta\tau_N)\leqslant\Phi(-2\delta\sqrt{n})\equiv\gamma_n^{(2)},$$

čia $\gamma_n^{(2)}\approx\Phi(-0{,}4)\approx 0{,}3446$, kai $n=100$, $\gamma_n^{(2)}\approx\Phi(-0{,}8944)\approx 0{,}1855$, kai $n=500$, ir $\gamma_n^{(2)}\approx\Phi(-1{,}7889)\approx 0{,}0368$, kai $n=2000$.

Normaliosios aproksimacijos paklaida Δ_n, remiantis Berry–Esseeno nelygybe ir prielaida $p(1-p)\geqslant 0{,}05$, tenkina nelygybę

$$\Delta_n\leqslant\frac{c_0}{\sqrt{n}}\frac{1}{\sqrt{p(1-p)}}\leqslant 3{,}424\frac{1}{\sqrt{n}}.$$

Taigi $\Delta_n\leqslant 0{,}3424$, kai $n=100$, $\Delta_n\leqslant 0{,}1531$, kai $n=500$, ir $\Delta_n\leqslant 0{,}0766$, kai $n=2000$.

Eksperimento rezultatai pateikiami lentelėje.

Lygis	Posekis	GSP		$n=100$		$n=500$		$n=2000$	
		Tikrasis dažnis	Prieskyra	Empirinis dažnis	Prieskyra	Empirinis dažnis	Prieskyra	Empirinis dažnis	Prieskyra
1	A	0.484	Dažnas	0.42	Dažnas	0.49	Dažnas	0.47	Dažnas
1	B	0.474	Dažnas	0.54	Dažnas	0.48	Dažnas	0.48	Dažnas
1	C	0.042	Retas	0.04	Retas	0.03	Retas	0.05	Retas
2	AA	0.224	Dažnas	0.20	Tarpinis	0.20	Tarpinis	0.21	Tarpinis
2	AB	0.240	Dažnas	0.20	Tarpinis	0.27	Dažnas	0.25	Dažnas
2	BA	0.240	Dažnas	0.31	Dažnas	0.23	Dažnas	0.24	Dažnas
2	BB	0.215	Dažnas	0.18	Retas	0.24	Dažnas	0.22	Dažnas
3	AAA	0.102	Retas	0.10	Retas	0.09	Retas	0.09	Retas
3	AAB	0.113	Retas	0.09	Retas	0.11	Retas	0.11	Retas
3	ABA	0.119	Retas	0.07	Retas	0.14	Retas	0.13	Retas
3	ABB	0.112	Retas	Netiriama	Netiriama	0.11	Retas	0.11	Retas
3	BAA	0.113	Retas	0.14	Retas	0.12	Retas	0.11	Retas
3	BAB	0.116	Retas	0.15	Retas	0.10	Retas	0.11	Retas
3	BBA	0.111	Retas	Netiriama	Netiriama	0.14	Retas	0.12	Retas
3	BBB	0.095	Retas	Netiriama	Netiriama	0.09	Retas	0.10	Retas

Išvados

Straipsnyje pasiūlytas tikimybinis dažnų posekių paieškos algoritmas, kuris analizuoja tam tikru būdu sudarytą pradinės duomenų bazės atsitiktinę imtį, ir remiantis šia analize daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje. Tikimybinis algoritmas nėra tikslus, tačiau jis veikia daug greičiau negu tikslieji algoritmai ir tinka žvalgomajai statistinei analizei. Tikimybinio algoritmo klaidų tikimybės įvertinamos statistiniais metodais. Didinant atsitiktinę imtį, klaidų tikimybė mažėja.

Literatūra

[1] J. Ayres, J. Flannick, J. Gehrke and T. Yiu. 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, 2002.

[2] V. Bagdonavičius, J. Kruopis. *Matematinė statika.* I dalis. TEV, Vilnius, 2007.

[3] J. Han, J. Pei, B. Mortazavi-asl, Q. Chen, U. Dayal and M.-Ch. Hsu. Freespan: Frequent pattern-projected sequential pattern mining. In *Proc. Knowledge Discovery and Data Mining*, pp. 355–359, 2000.

[4] J. Kubilius. *Tikimybių teorija ir matematinė statistika.* Mokslas, Vilnius, 1980.

[5] H.-Ch. Kum (Monica), J. Pei, W. Wang and D. Duncan. Approxmap: Approximate mining of consensus sequential patterns. In *Proceedings of the 2003 SIAM International Conference on Data Mining* (*SIAM DM '03*), pp. 311–315, 2003.

[6] *Online Trading Platform MetaTrader 4 History Center*. Available from Internet: http://www.metaquotes.net/data_center.

[7] J. Pei, J. Han, B. Mortazavi-asl, H. Pinto, Q. Chen, U. Dayal and M.-Ch. Hsu. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth. In *Proc. 17th International Conference on Data Engineering ICDE2001*, pp. 215–224, 2001.

[8] R. Srikant and R. Agrawal. Mining sequential patterns. In *Proceedings ICDE'95. Taipei* (*Taiwan*), 1995.

[9] R. Srikant and R. Agrawal. *Mining Sequential Patterns: Generalizations and Performance Improvements*. IBM Almaden Research Center, 1995.

[10] R. Tumasonis and G. Dzemyda. The probabilistic algorihm for mining frequent sequences. In *Proceedings ADBIS'04 Eight East-European Conference on Advances in Databases and Information Systems*, pp. 89–98, 2004.

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

SUMMARY

Probabilistic algorithm for mining frequent sequences

J. Pragarauskaitė and G. Dzemyda

Frequent sequence mining in large volume databases is important in many areas, e.g., biological, climate, financial databases. Exact frequent sequence mining algorithms usually read the whole database many times, and if the database is large enough, then frequent sequence mining is very long or requires supercomputers. A new probabilistic algorithm for mining frequent sequences is proposed. It analyzes a random sample of the initial database. The algorithm makes decisions about the initial database according to the random sample analysis results and performs much faster than the exact mining algorithms. The probability of errors made by the probabilistic algorithm is estimated using statistical methods.

Keywords: frequent sequence mining, probabilistic algorithm, data mining.