Tikimybinis dažnų posekių paieškos algoritmas

Julija Pragarauskaitė

Matematikos ir informatikos instituto doktorantė
Institute of Mathematics and Informatics,
Doctoral student
Akademijos g. 4, LT-08663 Vilnius
El. paštas: julija.pragarauskaite@gmail.com

Gintautas Dzemyda

Matematikos ir informatikos instituto profesorius
Institute of Mathematics and Informatics,
Professor
Akademijos g. 4, LT-08663 Vilnius
El. paštas: dzemyda@ktl.mii.lt

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 dažnų posekių 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. Tikimybinio algoritmo klaidų tikimybės įvertinamos statistiniais metodais. Tikimybinis algoritmas gali būti derinamas su tiksliaisiais dažnų posekių paieškos algoritmais. Jį galima taikyti ir bendrajam struktūrų paieškos uždaviniui.

Įvadas

Dažnų posekių paieška didelėse duomenų bazėse svarbi daugelyje sričių, pavyzdžiui, dažnų posekių paieška biologinių, klimato, finansinių duomenų bazėse, tinklalapių apsilankymų ir pardavimų duomenų bazėse. Pastaraisiais metais pasiūlyta daug tiksliųjų algoritmų dažnų posekių analizei. Populiariausias tikslusis algoritmas GSP (*Generalized Sequential Pattern mining algorithm*) (Srikant, Agrawal, 1995a; Srikant, Agrawal, 1995b) daug kartų perrenka visą pradinę duomenų bazę, nustato, kurie posekiai yra reti, ir jų toliau netiria. Kiti populiarūs tikslieji algoritmai yra SPADE (Zaki, 2001), SPAM (Ayres, Flannick, Gehrke, Yiu, 2002), PrefixSpan (Pei et al., 2001), FreeSpan (Han et al., 2000).

Jeigu duomenų bazė didelė, tai dažnų posekių paieška naudojant tiksliuosius algoritmus yra lėta arba reikalingi superkompiuteriai. Kai kuriuose uždaviniuose, pavyzdžiui, dažnų posekių tinklalapių apsilankymų bei pardavimų duomenų bazėse 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 daug trumpesnę duomenų imtį. Remiantis šia analize, daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje.

ProMFS (Tumasonis, Dzemyda, 2004) yra vienas iš apytikslių tikimybinių dažnų posekių paieškos algoritmų. Šis algoritmas generuoja naują trumpesnę seką, remdamasis statistinėmis pagrindinės sekos charakteristikomis: elemento pasirodymo sekoje tikimybe, sąlygine tikimybe, kad vienas elementas eis po kito, bei atstumų vidurkiu tarp dviejų elementų pagrindinėje sekoje. ProMFS algoritmas, remdamasis GSP, nustato dažnus posekius naujojoje sekoje ir daro išvadas apie dažnus posekius pradinėje duomenų bazėje.

Kitas apytikslis dažnų posekių nustatymo algoritmas yra ApproxMAP (Kum et al., 2003). Pagrindinė ApproxMAP algoritmo idėja yra vietoje tikslių posekių paieškos rasti posekius, apytiksliai naudojamus daugelyje kitų posekių.

Straipsnyje pasiūlytas naujas tikimybinis dažnų posekių paieškos algoritmas analizuoja tam tikru būdu sudarytą pradinės duomenų bazės atsitiktinę imtį. Remiantis šia analize daromos statistinės išvados apie dažnus posekius pradinėje duomenų bazėje. Tikimybinio algoritmo klaidų tikimybės įvertinamos statistiniais metodais. Didinant atsitiktinę imtį, klaidų tikimybės mažėja. Pasiūlytą tikimybinį algoritmą galima derinti su tiksliaisiais dažnų posekių paieškos algoritmais. Jį galima taikyti ir bendrajam struktūrų paieškos (angl. *sequential pattern mining*) uždaviniui.

GSP agoritmas *(Generalized Sequential Pattern mining algorithm)*

GSP yra tikslus algoritmas dažniems posekiams nustatyti pagal pasirinktą dažnio slenkstį $\varepsilon \in (0,1)$. Tarkime, kad pradinė duomenų bazė, kurioje ieškosime dažnai pasikartojančių posekių, yra sunumeruota duomenų aibė $S = (S_1, S_2, \dots, S_n)$. Aibės S elementai S_i gali įgyti m skirtingų reikšmių $a_1, a_2, \dots, a_m$. Posekis $a_{i_1}, \dots, a_{i_k}$ vadinamas dažnu, jeigu

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

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 $a_1, a_2, \dots, a_m$ ir nustato, kurie posekiai yra dažni. Toliau iš nustatytų dažnų posekių formuojami antrojo lygio (dviejų simbolių) kandidatai $a_1a_1, a_1a_2, \dots, a_1a_n, a_2a_1, \dots, a_2a_n, \dots, a_na_1, \dots, a_na_n$, 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. Dviejų simbolių dažnų posekių nustatymas vykdomas dar kartą perrenkant visą pradinę duomenų bazę.

Panašiai nustatomi trečio ir kitų lygių dažni posekiai. Algoritmas baigiamas, kai eiliniame lygyje nebėra kandidatų į dažnus posekius.

Tikimybinis algoritmas

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. Siūlomas tikimybinis algoritmas yra daug spartesnis, 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 $1, 2, \dots, N$ su vienodomis tikimybėmis $\frac{1}{N}$, realizacijų seką $\eta_1, \eta_2, \dots, \eta_n$.
- Ieškant pirmojo lygio (vieno elemento) dažnų posekių, atsitiktinė imtis $\bar{S}$ elementams a_i yra tiesiog $S_{\eta_1}, S_{\eta_2}, \dots, S_{\eta_n}$. Antrojo lygio atsitiktinė imtis elementų poroms a_ia_j yra, $(S_{\eta_2}, S_{\eta_2+1})$, …, $(S_{\eta_n}, S_{\eta_n+1})$. *k*-ojo lygio atsitiktinė imtis elementų rinkiniams $a_i \dots a_k$ yra $(S_{\eta_1}, \dots, S_{\eta_1+k-1})$, $(S_{\eta_2}, \dots, S_{\eta_2+k-1})$, … , $(S_{\eta_n}, \dots, S_{\eta_n+k-1})$ ir t. t. Tokia imtis yra sudaryta grąžtinniu ėmimu, nes kai kurie skaičiai η_i gali pasikartoti. Negrąžtinniu 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 $\eta_1, \eta_2, \dots, \eta_n$.

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

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

Pasirenkame skaičių

$$\delta > 0\,(0 < \varepsilon - \delta < \varepsilon + \delta < 1), k = 1,2, \dots.$$

Posekius $a_{i_1}, a_{i_2}, \dots, a_{i_k}$ klasifikuojame į tris grupes:

1) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \geq \varepsilon + \delta$, tai posekį $a_{i_1}, \dots, a_{i_k}$ priskiriame dažnų posekių klasei;

2) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \leq \varepsilon - \delta$, tai posekį $a_{i_1}, \dots, a_{i_k}$ priskiriame retų posekių klasei;

3) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \in (\varepsilon - \delta, \varepsilon + \delta)$, tai posekį $a_{i_1}, \dots, a_{i_k}$ priskiriame tarpinių posekių klasei.

Aptarsime tikimybinio algoritmo klaidų tikimybių įvertinius. Fiksuokime kokį nors posekį $a_{i_1}, \dots, a_{i_k}$. 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 $\bar{p}_n = \bar{p}_n(a_{i_1}, \dots, a_{i_k})$, $p = p(a_{i_1}, \dots, a_{i_k})$. Akivaizdu, kad pirmosios rūšies klaidos tikimybė neviršija

$$\max_{p<\varepsilon} P(\overline{p_n} - p > \delta), \quad (1)$$

o antrosios rūšies klaidos tikimybė neviršija

$$\max_{p\geq\varepsilon} P(\overline{p_n} - p < -\delta). \quad (2)$$

Vertinant šias tikimybes, patogu pasinaudoti tokia schema. Apibrėžkime atsitiktinius dydžius

$$Z_i = \begin{cases} 1, \text{jeigu } S_{\eta_i} = a_{i_1}, S_{\eta_i+1} = a_{i_2}, \dots, S_{\eta_i+k-1} = a_{i_k} \\ 0, \quad \text{priešingu atveju} \end{cases},$$

čia $i = 1, \dots, n$.

Dėl sekos $\eta_1, \eta_2, \dots, \eta_n$ sudarymo būdo atsitiktiniai dydžiai $Z_1, Z_2, \dots, Z_n$ yra tarpusavyje nepriklausomi ir vienodai pasiskirtę (Bernulio eksperimentų schema). Matome, kad atsitiktinių dydžių Z_i vidurkis

$$EZ_i = p, \quad (3)$$

o dispersija

$$DZ_i = p(1-p). \quad (4)$$

Tikimybes (1) ir (2) galima įvertinti standartiniais matematinės statistikos metodais: remiantis binominio skirstinio savybėmis grąžintinės imties atveju (Bagdonavičius, Kruopis, 2007, p. 220–221) bei hipergeometrinio skirstinio savybėmis negrąžintinės imties atveju (Bagdonavičius, Kruopis, 2007, p. 224).

Apsiribosime asimptotiniais klaidų tikimybių įvertiniais grąžintinės imties atveju. Jie veiksmingi, kai imties didumas n yra pakankamai didelis. Apibrėžkime atsitiktinį dydį

$$\Sigma_n = Z_1 + Z_2 + \dots + Z_n.$$

Remiantis centrine ribine teorema (Kubilius, 1980, p. 264–271), su visais $a \leq b$

$$P\left(a \leq \frac{\Sigma_n - E\Sigma_n}{\sqrt{D\Sigma_n}} \leq b\right) \to \Phi(b) - \Phi(a), n \to \infty;$$

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

Kadangi $\Sigma_n = \bar{p}_n n$, $E\Sigma_n = np$ ir $D\Sigma_n = np(1-p)$, tai su visais $a \leq b$

$$P(a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \overline{p_n} - p \leq b\frac{\sqrt{p(1-p)}}{\sqrt{n}}) \to$$

$$\to \Phi(b) - \Phi(a), n \to \infty.$$

Jeigu $a = -\infty$, tai su visais b

$$P(\overline{p_n} - p \leq b\frac{\sqrt{p(1-p)}}{\sqrt{n}}) \to \Phi(b), n \to \infty. \quad (5)$$

Jeigu $b = +\infty$, tai su visais a

$$P\left(a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \overline{p_n} - p\right) \to 1 - \Phi(a), n \to \infty. \quad (6)$$

Jeigu n pakankamai didelis, tai, remiantis (5) ir (6),

$$\max_{p<\varepsilon} P(\overline{p_n} - p > \delta) \approx$$

$$\approx \max_{p<\varepsilon}\left(1 - \Phi(\delta\frac{\sqrt{n}}{\sqrt{p(1-p)}})\right) \leq 1 - \Phi(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_0(1-\varepsilon_0)}}), \quad (7)$$

čia $\varepsilon_0 = \min\left(\varepsilon, \frac{1}{2}\right)$, ir

$$\max_{p\geq\varepsilon} P(\overline{p_n} - p < -\delta) \approx$$

$$\approx \max_{p\geq\varepsilon}\left(\Phi(-\delta\frac{\sqrt{n}}{\sqrt{p(1-p)}})\right) \leq \Phi(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}), \quad (8)$$

čia $\varepsilon_1 = \max\left(\frac{1}{2}, \varepsilon\right)$.

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

$$P(\bar{p}_n \geq \varepsilon) \to \frac{1}{2}, n \to \infty$$

ir

$$P(\bar{p}_n < \varepsilon) \to \frac{1}{2}, n \to \infty.$$

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, jis yra artimas empiriniam dažniui $\bar{p}_n$, kai n yra pakankamai didelis, nes vėl remiantis centrine ribine teorema su visais $\mu > 0$

$$P(|\bar{p}_n - p| > \mu) \to 0, n \to \infty.$$

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 $\delta\sqrt{n} = \text{const}$.

Pavyzdys. Tarkime, kad $\delta = 0{,}05;\ \varepsilon = 0{,}2;\ n = 100$. Remiantis (7), klaidos, jog posekis priskirtas dažnų posekių klasei, nors iš tikro jis yra retas, tikimybė neviršija

$$1 - \Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(1{,}25) \approx 0{,}11.$$

Remiantis (8), klaidos, jog posekis priskirtas retų posekių klasei, nors iš tikro jis yra dažnas, tikimybė neviršija

$$\Phi\left(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-1{,}0) \approx 0{,}1587.$$

Tarkime, kad $\delta = 0{,}05;\ \varepsilon = 0{,}2;\ n = 400$. Remiantis (7), klaidos, jog posekis priskirtas dažnų posekių klasei, nors iš tikro jis yra retas, tikimybė neviršija

$$1 - \Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(2{,}5) \approx 0{,}006.$$

Remiantis (8), klaidos, jog posekis priskirtas retų posekių klasei, nors iš tikro jis yra dažnas, tikimybė neviršija

$$\Phi\left(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-2{,}0) \approx 0{,}0228.$$

Eksperimentas

Tirsime finansinių duomenų bazę – valiutų EUR–USD poros valandinius duomenis nuo 1999 m. sausio 4 dienos 10:00 val. iki 2009 m. gegužės 18 dienos 16:00 valandos (duomenys paimti iš *Online Trading Platform MetaTrader 4 History Center*). Finansinės duomenų bazės $S = (S_1, S_2, \ldots, S_N)$ elementų skaičius $N = 64467$, o jos elementai gali įgyti šias skirtingas reikšmes $\{A, B, C\}$:

- A – jeigu *i*-osios valandos pabaigos kursas C_i yra didesnis nei valandos pradžios kursas O_i, t. y. $S_i = A$, jeigu $C_i > O_i$;
- B – jeigu *i*-osios valandos pabaigos kursas C_i yra mažesnis nei valandos pradžios kursas O_i, t. y. $S_i = B$, jeigu $C_i < O_i$;
- C – jeigu *i*-osios valandos pabaigos kursas C_i yra lygus valandos pradžios kursui O_i, t. y. $S_i = C$, jeigu $C_i = O_i$.

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 $0{,}08$, t. y. $\varepsilon = 0{,}08$.

Pirmiausia ištiriame pradinę seką su GSP algoritmu ir nustatome dažnus posekius. Tada taikome tikimybinį algoritmą dažnų posekių paieškai pradinėje duomenų bazėje. Pasirenkame atsitiktinės imties didumą $n = 100$, $n = 500$ ir $n = 2000$ bei $\delta = 0{,}02$.

Remiantis (7), pirmosios rūšies klaidos (posekis priskirtas dažnų posekių klasei, nors iš tikro jis yra retas) tikimybės įverčiai tokie:

$n = 100$:

$$1 - \Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(0{,}7372) \approx 0{,}2305;$$

$n = 500$:

$$1 - \Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(1{,}6485) \approx 0{,}0496;$$

$n = 2000$:

$$1 - \Phi\left(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(3{,}2969) \approx 0{,}0005.$$

$n = 100$:

$$\Phi\left(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-0{,}4) \approx 0{,}3446;$$

$n = 500$:

$$\Phi\left(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-0{,}8944) \approx 0{,}1855;$$

$n = 2000$:

$$\Phi\left(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-1{,}7889) \approx 0{,}0368.$$

Remiantis (8), antrosios rūšies klaidos (posekis priskirtas retų posekių klasei, nors iš tikro jis yra dažnas) tikimybės įverčiai tokie:

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	0,482	Dažnas	0,39	Dažnas	0,494	Dažnas	0,488	Dažnas
1	B	0,472	Dažnas	0,56	Dažnas	0,472	Dažnas	0,4675	Dažnas
1	C	0,046	Retas	0,05	Retas	0,034	Retas	0,0445	Retas
2	AA	0,222	Dažnas	0,14	Dažnas	0,224	Dažnas	0,213	Dažnas
2	AB	0,236	Dažnas	0,24	Dažnas	0,244	Dažnas	0,2535	Dažnas
2	BA	0,237	Dažnas	0,31	Dažnas	0,236	Dažnas	0,238	Dažnas
2	BB	0,214	Dažnas	0,23	Dažnas	0,214	Dažnas	0,204	Dažnas
3	AAA	0,101	Dažnas	0,03	Retas	0,088	Tarpinis	0,102	Dažnas
3	AAB	0,112	Dažnas	0,1	Dažnas	0,13	Dažnas	0,1	Dažnas
3	ABA	0,116	Dažnas	0,12	Dažnas	0,11	Dažna	0,1255	Dažnas
3	ABB	0,110	Dažnas	0,11	Dažnas	0,126	Dažnas	0,1145	Dažnas
3	BAA	0,112	Dažnas	0,13	Dažnas	0,092	Tarpinis	0,122	Dažnas
3	BAB	0,114	Dažnas	0,18	Dažnas	0,114	Dažnas	0,104	Dažnas
3	BBA	0,110	Dažnas	0,16	Dažnas	0,112	Dažnas	0,1085	Dažnas
3	BBB	0,095	Dažnas	0,07	Tarpinis	0,09	Tarpinis	0,0865	Tarpinis
4	AAAA	0,044	Retas	0,01	Tarpinis	0,032	Retas	0,048	Retas
4	AAAB	0,052	Retas	0,02	Retas	0,054	Retas	0,049	Retas
4	AABA	0,054	Retas	0,06	Retas	0,066	Tarpinis	0,041	Retas
4	AABB	0,056	Retas	0,04	Retas	0,056	Retas	0,054	Retas
4	ABAA	0,054	Retas	0,08	Tarpinis	0,052	Retas	0,0565	Retas
4	ABAB	0,056	Retas	0,03	Retas	0,05	Retas	0,0615	Tarpinis
4	ABBA	0,055	Retas	0,02	Tarpinis	0,054	Retas	0,051	Retas
4	ABBB	0,050	Retas	0,07	Tarpinis	0,068	Tarpinis	0,0535	Retas
4	BAAA	0,059	Retas	0,08	Tarpinis	0,05	Retas	0,055	Retas
4	BAAB	0,055	Retas	0,05	Retas	0,04	Retas	0,0605	Tarpinis
4	BABA	0,057	Retas	0,09	Dažnis	0,052	Retas	0,0525	Retas
4	BABB	0,052	Retas	0,07	Tarpinis	0,058	Retas	0,0475	Retas
4	BBAA	0,052	Retas	0,07	Tarpinis	0,042	Retas	0,053	Retas
4	BBAB	0,052	Retas	0,08	Tarpinis	0,064	Tarpinis	0,052	Retas
4	BBBA	0,050	Retas	0,04	Retas	0,058	Retas	0,041	Retas
4	BBBB	0,041	Retas	0,03	Retas	0,03	Retas	0,0415	Retas

Išvados

Straipsnyje pasiūlytas naujas 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.

Eksperimento rezultatai parodė, kad pasiūlytas tikimybinis algoritmas, analizuodamas atsitiktinę imtį, sudarytą tik iš 100 simbolių, kai pradinė duomenų seka turi 64 467 simbolius, sugeba klasifikuoti posekius į dažnus ir retus esant klaidų tikimybei 0,2305 (pirmosios rūšies klaida) ir 0,3446 (antrosios rūšies klaida). Didinant atsitiktinę imtį iki 2000 simbolių, klaidų tikimybės mažėja iki 0,0005 (pirmosios rūšies klaida) ir 0,0368 (antrosios rūšies klaida).

Tikimybinis algoritmas gali būti derinamas su tiksliaisiais dažnų posekių paieškos algoritmais. Jį galima taikyti ir bendrajam struktūrų paieškos uždaviniui.

LITERATŪRA

AYRES, Jay; FLANNICK, Jason; GEHRKE, Johannes; YIU, Tomi (2002). Sequential Pattern mining using a bitmap representation. Iš *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data minin*g, p. 429–435.

BAGDONAVIČIUS, Vilijandas; KRUOPIS, Julius (2007). *Matematinė statika.* I dalis. Vilnius: TEV.

HAN, Jiawei; PEI, Jian; MORTAZAVI-ASL, Behzad; CHEN, Qiming; DAYAL, Umeshwar; HSU, Mei-Chun (2000). FreeSpan: Frequent pattern-projected sequential pattern mining. Iš *Proc. Knowledge Discovery and Data Mining*, p. 355–359.

KUBILIUS, Jonas (1980). *Tikimybių teorija ir matematinė statistika*. Vilnius: Mokslas.

KUM, Hye-Chung (Monica); PEI, Jian; WANG, Wei; DUNCAN, Dean (2003). ApproxMAP: Approximate Mining of Consensus Sequential Patterns. Iš *Proceedings of the 2003 SIAM International Conference on Data Mining* (SIAM DM '03), p. 311–315.

Online Trading Platform MetaTrader 4 History Center [žiūrėta 2009 m. gegužės 22 d]. Prieiga per internetą: <http://www.metaquotes.net/data_center>.

PEI, Jian; HAN, Jiawei; MORTAZAVI-ASL, Behzad; PINTO, Helen; CHEN, Qimin; DAYAL, Umeshwar; HSU, Mei-Chun (2001). PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. Iš *Proc. 17th International Conference on Data Engineering ICDE2001*, p. 215–224.

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh (1995). Mining sequential patterns. Iš *Proceedings ICDE'95. Taipei (Taiwan).*

SRIKANT, Ramakrishnan; AGRAWAL, Rakesh (1995). *Mining Sequential Patterns: Generalizations and Performance Improvements*. IBM Almaden Research Center.

TUMASONIS, Romanas; DZEMYDA, Gintautas (2004). The Probabilistic Algorihm for Mining Frequent Sequences. Iš *Proceedings ADBIS'04 Eight East-European Conference on Advances in Databases and Information Systems*, p. 89–98.

ZAKI, Mohammed J. (2001). SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning Journal*, vol. 42, no. 1–2, p. 31–60.

PROBABILISTIC ALGORITHM FOR MINING FREQUENT SEQUENCES

Julija Pragarauskaitė, Gintautas Dzemyda

Summary

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. The algorithm can be used together with the exact frequent sequence mining algorithms.