VILNIAUS UNIVERSITETAS

JULIJA PRAGARAUSKAITĖ

DAŽNŲ SEKŲ ANALIZĖ SPRENDIMŲ PRIĖMIMUI LABAI DIDELĖSE DUOMENŲ BAZĖSE

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09 P)

Vilnius, 2013

Disertacija rengta 2008 – 2013 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinis vadovas:

prof. habil. dr. Gintautas Dzemyda (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Nariai:

prof. habil. dr. Juozas Augutis (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika – 09 P),

prof. dr. Albertas Čaplinskas (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P),

prof. habil. dr. Antanas Čenys (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T),

prof. habil. dr. Mifodijus Sapagovas (Vilniaus universitetas, fiziniai mokslai, matematika – 01 P).

Oponentai:

prof. dr. Eduardas Bareiša (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

doc. dr. Olga Kurasova (Vilniaus universitetas, fiziniai mokslai, informatika – 09 P).

Disertacija bus ginama Vilniaus universiteto viešame Informatikos mokslo krypties tarybos posėdyje 2013 m. birželio 27 d. 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje. Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2013 m. gegužės 27 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

JULIJA PRAGARAUSKAITĖ

FREQUENT PATTERN ANALYSIS FOR DECISION MAKING IN BIG DATA

Summary of Doctoral Dissertation
Physical Sciences, Informatics (09 P)

Vilnius, 2013

Doctoral dissertation was prepared at Institute of Mathematics and Informatics of Vilnius University in 2008 – 2013.

Scientific Supervisor:

Prof. Dr. Habil. Gintautas Dzemyda (Vilnius University, Physical Sciences, Informatics – 09 P).

This dissertation will be defended at the Council of the Scientific Field of Informatics of Vilnius University:

Chairman:

Prof. Dr. Romas Baronas (Vilnius University, Physical Sciences, Informatics – 09 P).

Members:

Prof. Dr. Habil. Juozas Augutis (Vytautas Magnus University, Physical Sciences, Informatics – 09 P),

Prof. Dr. Albertas Čaplinskas (Vilnius University, Physical Sciences, Informatics – 09 P),

Prof. Dr. Habil. Antanas Čenys (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T),

Prof. Dr. Habil. Mifodijus Sapagovas (Vilnius University, Physical Sciences, Mathematics – 01 P).

Opponents:

Prof. Dr. Eduardas Bareiša (Kaunas University of Technology, Technological Sciences, Informatics Engineering – 07 T),

Assoc. Dr. Olga Kurasova (Vilnius University, Physical Sciences, Informatics – 09 P).

This dissertation will be defended at the public meeting of the Council of the Scientific Fielg of Informatics in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p.m. on the 27th of June 2013. Address: Akademijos st. 4, LT-08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed on the 27th of May 2013.

A copy of the doctoral dissertation is available for review at the Library of Vilnius University.

Trumpiniai, naudojami šioje disertacijos santraukoje

ApproxMAP	Approximate Multiple Alignment Pattern mining algorithm
GSP	Generalized Sequence Patterns algorithm
MDS	Multidimensional scaling
LAPIN	Last Position Induction algorithm
MPBM	Markov Property Based Method
MRM	Multiple Re-sampling Method
PRISM	Prime-Encoding Based Sequence Mining algorithm
ProMFS	Probabilistic algorithm for Mining Frequent Sequences
RSM	Random Sampling Method
SOM	Self Organizing Map
SPADE	Sequential Pattern Discovery using equivalence classes algorithm
SPAM	Sequential Pattern Mining algorithm

Įvadas

Tyrimų sritis

Kiekvieną dieną pasaulyje sukaupiami didžiuliai informacijos kiekiai ir jie sparčiai auga. Sukaupti informacijos kiekiai leidžia atskleisti įvairią informaciją apie sukauptus duomenis: daryti finansines prognozes, nustatyti įvairias ligas, identifikuoti potencialius nusikaltimus ir t. t. Pasak Moore dėsnio, kompiuterių galingumas padvigubėja apytiksliai kas 18 mėnesių, o tuo tarpu sukaupta informacija pasaulyje padvigubėja apytiksliai kas 40 mėnesių, o nuo 2012 m. kiekvieną dieną sukaupiamas papildomas kvintilijonas baitų (IBM review on what is big data, retrieved 2013). Toks spartus kompiuterių galios ir

sukaupiamų duomenų kiekio augimas bei šiuolaikiniai algoritmai gali atskleisti svarbius faktus bei įžvalgas, kurios anksčiau nebuvo pastebėtos. Apytiksliai duomenų tyrybos algoritmai yra labai svarbūs analizuojant tokius didelius duomenų kiekius, nes algoritmų greitis yra ypač svarbus daugelyje sričių, tuo tarpu tikslieji metodai paprastai yra lėti bei naudojami tik uždaviniuose, kuriuose reikalingas tikslus atsakymas (Han, Kamber, 2006). Terminas „didelės duomenų bazės“ (angl. *big data*) yra dažnai naudojamas duomenų tyryboje pastaruoju metu. Jis aprašo duomenų aibę, kuri yra tokia didelė, kad ją sunkiai įmanoma apdoroti naudojant standartines duomenų bazių valdymo sistemas ar tradicinius algoritmus. Problemos didelėse duomenų bazėse iškyla kaupiant duomenis, vykdant paiešką, perkeliant duomenis iš vienos vietos į kitą, analizuojant arba vizualizuojant duomenis. Duomenų analitikai dažnai susiduria su šiomis problemomis analizuojant dideles duomenų bazes įvairiose srityse, pvz. meteorologijoje, genetikoje (Editorial review in Nature, 2008), fizikinėse simuliacijose, biologiniuose duomenyse, interneto duomenyse, finansuose, marketinge ir t. t.

Tikslieji dažnų sekų paieškos algoritmai tiksliai nustato dažnas ir retas sekas pradinėje duomenų bazėje, tačiau jie kelis kartus skaito pradinę duomenų bazę ir didelėse duomenų bazėse tai tampa ilga bei brangia užduotimi. Apytiksliai dažnų sekų paieškos metodai su įvertinta paklaidų tikimybe yra tinkami naudoti daugelyje sričių ir taikymų, kuriuose algoritmo tikslumas nėra toks svarbus kaip greitis, pvz. marketinge, interneto vartotojų analizėje, finansuose, valiutų ir akcijų kursuose, ir t. t. Apytiksliai metodai yra daug greitesni nei tikslieji, nes užuot kelis kartus skaitę pradinę duomenų bazę, jie analizuoja gerokai trumpesnę pradinės duomenų bazės imtį bei daro išvadas apie dažnas sekas pradinėje duomenų bazėje.

Mūsų žiniomis, visi anksčiau pasiūlyti apytiksliai dažnų sekų paieškos metodai neturi teorinio metodo daromų klaidų įvertinimo, tačiau remiasi išplėstiniais empiriniais tyrimais ir stebėjimais naudojant įvairias duomenų bazes, kad būtų galima įvertinti metodo tikslumą bei greitį.

Šioje disertacijoje yra pasiūlyti nauji apytiksliai dažnų sekų paieškos metodai su teoriniu metodo daromų paklaidų įvertinimu, kuomet sekos yra klasifikuojamos į dažnas bei retas. Metodų rezultatai buvo palyginti su kitais tiksliaisiais bei apytiksliais dažnų sekų

paieškos algoritmais ir pateiktos rekomendacijos parametrų parinkimui pasiūlytuose apytiksliuose metoduose.

Darbo tikslai ir uždaviniai

Šios disertacijos tikslai yra:

(1) sukurti naujus apytikslius dažnų sekų paieškos metodus su įvertintomis metodų daromomis paklaidų tikimybėmis klasifikuojant sekas į dažnas ir retas;
(2) pasiūlyti metodologiją duomenų vizualizavimui didelėse duomenų bazėse.

Šiems tikslams pasiekti buvo iškelti tokie uždaviniai:

- Ištirti egzistuojančius dažnų sekų paieškos algoritmus (tiksliuosius ir apytikslius);
- Pasiūlyti apytikslius dažnų sekų paieškos metodus, kurie turėtų įvertintas klasifikavimo į dažnas ir retas sekas klaidų tikimybes;
- Pasiūlyti atitiktinės imties metodą dažnų sekų paieškai su įvertintomis metodo daromomis klaidų tikimybėmis klasifikuojant sekas į dažnas ir retas;
- Pasiūlyti apytikslį dažnų sekų paieškos metodą, kuris remiasi Markovo savybe;
- Įvertinti pasiūlytų metodų tikslumą ir greitį bei palyginti su kitais tiksliaisiais ir apytiksliais algoritmais;
- Pasiūlyti metodologiją interneto vartotojų elgsenos analizei ir vizualizacijai.

Darbo mokslinis naujumas

Kaip sprendimas anksčiau įvardintai dažnų sekų paieškos problemai didelėse duomenų bazėse šioje disertacijoje buvo pasiūlyti trys nauji apytiksliai dažnų sekų paieškos metodai bei metodologija, skirta interneto vartotojų elgsenos analizei bei vizualizavimui. Pasiūlyti apytiksliai metodai buvo testuojami naudojant tikrą bei dirbtinai sugeneruotą duomenų bazes:

- Atsitiktinės imties metodas (Random Sampling Method - RSM) formuoja pradinės duomenų bazės atsitiktinę imtį ir nustato dažnas sekas remiantis

atsitiktinės imties analizės rezultatais. Šio metodo privalumas – teorinis paklaidų tikimybių įvertinimas naudojantis standartiniais statistiniais metodais.

- Daugybinio perskaičiavimo metodas (Multiple Re-sampling Method – MRM) – RSM metodo patobulinimas, kuris formuoja kelias pradinės duomenų bazės atsitiktines imtis ir taip sumažina paklaidų tikimybes.
- Markovo savybe besiremiantis metodas (Markov Property Based Method – MPBM) kelis kartus skaito pradinę duomenų bazę, priklausomai nuo Markovo proceso eilės, bei apskaičiuoja empirinius dažnius remdamasis Markovo savybe.

Didelių duomenų vizualizavimui buvo naudojami pirkėjų internetu elgsenos duomenys, kurie buvo analizuojami naudojant geometrinius metodus, daugiamates skales bei neuroninius tinklus. Saviorganizuojantis neuroninis tinklas pademonstravo geriausius rezultatus, tačiau jis reikalauja išplėstinių žinių apie duomenis ir parametrus.

Ginamieji disertacijos teiginiai

1) Apytiksliai (tikimybiniai) dažnų sekų paieškos metodai gali duoti greitus rezultatus, kas yra labai svarbu taip sparčiai augant informacijos kiekiui pasaulyje. Tuo tarpu tikslieji dažnų sekų paieškos metodai grąžina tikslų atsakymą, kurio sekos yra dažnos, tačiau jų greitis neatitinka reikalavimų daugelyje taikymų ir sričių, kuriose šie metodai yra naudojami.
2) Atsitiktinės imties metodas (RSM) turi didelį privalumą, nes jame yra teoriškai įvertintos metodo daromos paklaidų tikimybės naudojantis standartiniais statistiniais metodais. Teorinis įvertinimas leidžia naudoti metodą be išplėstinių eksperimentinių tyrimų, kurie yra būtini kituose dažnų sekų paieškos algoritmuose.
3) Daugybinio perskaičiavimo metodas (MRM) yra RSM metodo patobulinimas, kuris formuoja kelias pradinės duomenų bazės atsitiktines imtis ir taip sumažina paklaidų tikimybes. Jis eliminuoja tarpinių sekų klasę iš RSM metodo taip pagerindamas MRM metodo tikslumą, tačiau yra lėtesnis už RSM tiek kartų, kiek papildomų atsitiktinių imčių yra analizuojama.

4) Apytiksliai dažnų sekų paieškos metodai, kurie visiškai nuskaito pradinę duomenų bazę, netenkina greičio reikalavimų, kurie yra labai svarbūs daugelyje taikymų ir sričių. Šioje disertacijoje pasiūlytas apytikslis Markovo savybe besiremiantis metodas (MPBM) netenkina greičio reikalavimų, nes kelis kartus skaito pradinę duomenų bazę, priklausomai nuo Markovo proceso eilės, bei apskaičiuoja empirinius dažnius remdamasis Markovo savybe.
5) Interneto vartotojų elgsenos analizėje bei vizualizavime saviorganizuojantis neuroninis tinklas pademonstravo geriausius rezultatus, tačiau jis reikalauja išplėstinių žinių apie duomenis ir parametrus.

Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 5 moksliniuose leidiniuose: 4 periodiniuose recenzuojamose mokslo žurnaluose ir 1 konferencijos pranešimų medžiagoje.

Disertacijos struktūra

Disertaciją sudaro įvadas, 4 skyriai ir išvados. Pirmajame skyriuje pateikiama dažnų sekų apžvalga bei apibrėžimas. Antrajame skyriuje pateikiama išsami informacija apie tiksliuosius dažnų sekų paieškos algoritmus bei ju veikimo principus. Trečiajame skyriuje pateikiama detali informacija apie apytikslius egzistuojančius dažnų sekų algoritmus bei šioje disertacijoje pasiūlytus apytikslius dažnų sekų paieškos metodus. Ketvirtajame skyriuje pateikiami disertacijoje pasiūlytų metodų eksperimentinių tyrimų rezultatai. Ketvirtajame skyriuje pateikiama metadologija, skirta internetinių vartotojų elgsenos analizei bei vizualizavimui.

1. Dažnų sekų paieškos algoritmų apžvalga ir problemos

1.1. Įžanga

Dažnų sekų paieška pirmą kartą pritaikyta 1993 m. pirkėjo krepšelio analizei (Agrawal et al., 1993) ir nuo to laiko dažnų sekų paieška tapo plačiai nagrinėjama sritimi. Visi dažnų sekų paieškos algoritmai gali būti klasifikuojami į tiksliuosius bei apytikslius (tikimybinius) algoritmus:

- Tikslieji dažnų sekų paieškos algoritmai kelis kartus skaito pradinę duomenų bazę ir jei duomenų bazė yra didelė, tuomet tai gali užtrukti gana ilgai ir tapti brangia užduotimi, kuri naudoja daug kompiuterio resursų.
- Apytiksliai dažnų sekų paieškos algoritmai yra gerokai greitesni nei tikslieji metodai, nes užuot nuodugniai skaitę pradinę duomenų bazę, jie analizuoja daug mažesnę, specifiniu būdu sudarytą pradinės duomenų bazės imtį bei daro sprendimus apie dažnas ir retas sekas.

Ir tikslieji, ir apytiksliai algoritmai yra nepakeičiami daugelyje sričių bei taikymų, o tinkamiausias algoritmas yra pasirenkamas pagal reikalavimus, kurie nurodo, kas yra svarbiau: ar algoritmo greitis, ar tikslumas. Tikslieji algortimai yra labai svarbūs medicinoje, genetikoje, biologijoje bei kitose srityse, kur reikalaujamas tikslus atsakymas; tuo tarpu apytiksliai algortimai greitai grąžina apytikslius rezultatus bei yra tinkami daugelyje sričių, kur algoritmo paklaidos yra priimtinos, pvz. interneto vartotojų elgsenos analizėje, marketinge, finansinuose duomenyse ir t. t. Taip sparčiai augant duomenų kiekiams pasaulyje tapo aišku, kad apytikslių dažnų sekų paieškos algoritmų pritaikomumas tik didės, o teorinis metodo daromų paklaidų įvertinimas leidžia dar lanksčiau panaudoti apytikslius metodus, nes jų daromos paklaidos būna žinomos iš anksto. Apytiksliai dažnų sekų paieškos metodai taip pat yra nepakeičiami begaliniuose duomenų srautuose, kurie yra nuolat papildomi naujais duomenimis, ir tokių srautų analizėje pilnas duomenų skaitymas yra neįmanomas.

Dažnų sekų paieškos studijos gali būti suskirstytos į dvi kategorijas: dažnų sekų paieška (angl. *frequent sequence*) bei dažnų rinkinių paieška (angl. *frequent itemset*) (Han et al.,

2007). Dažnų rinkinių paieška susitelkia ties dažnų rinkinių dažnių nustatymu, kur elementų tvarka rinkiniuose nėra svarbi (Agrawal, Srikant, 1994; Brin et al., 1997; Han et al., 2007; Han et al., 2000, Park et al., 1995; Park et al., 1995; Sarawagi et al., 2000; Savasere et al., 1995; Zaki, 2000). Dažnų rinkinių paieška dažnai vadinama susietumo taisyklių (angl. *association rules*) paieška pradinėje duomenų bazėje. Dažnų sekų paieškoje, priešingai nei dažnų rinkinių paieškoje, elementų tvarka yra labai svarbi (Agrawal, Srikant, 1995; Ayres et al., 2002; Han et al., 2001; Srikant, Agrawal, 1996; Zaki, 2001).

1.2. Definition of the frequent sequence

Laikykime, kad pradinė duomenų bazė yra $S = \{x_1, x_2, \dots, x_N\}$, kurios elementai x_i gali turėti M skirtingų reikšmių iš aibės $V = \{\alpha_1, \alpha_2, \dots, \alpha_M\}$. Seka $(\alpha_{i_1}, \dots, \alpha_{i_m})$ yra laikoma dažna, jeigu (# yra skaičius):

$$p(\alpha_{i_1}, \dots, \alpha_{i_k}) = \frac{1}{N-m+1} \#\{j \in \{1, \dots, N-m+1\}: x_j = \alpha_{i_1}, x_{j+1} = \alpha_{i_2}, \dots, x_{j+m-1} = \alpha_{i_m}\} \geq \varepsilon,$$

kur $\varepsilon \, \epsilon \, (0,1)$ yra vartotojo pasirinktas dažnio slenkstis (angl. *minimum support threshold*).

2. Tikslieji dažnų sekų paieškos algoritmai

Tikslieji dažnų sekų paieškos algoritmai tiksliai nustato dažnas ir retas sekas bei jų dažnius pradinėje duomenų bazėje, tačiau jie skaito pradinę duomenų bazę bent kelis kartus. Labai didelėse duomenų bazėse tikslieji dažnų sekų paieškos algoritmai gali būti labai lėti ir netgi sunkiai panaudojami tam tikrose srityse, kur algoritmų greitis yra daug svarbiau palyginus su algoritmo tikslumu, pvz. marketinge, interneto vartotojų analizėje, akcijų bei valiutų kursų duomenyse, biologinėse duomenų bazėse ir t. t. Akcijų ar valiutų kursuose net mažiausias uždelsimas identifikuojant dažnas ir retas sekas gali paveikti, kad tinkamiausias momentas sprendimui dėl investicijos jau gali būti praėjęs.

GSP (angl. *Generalized Sequence Patterns*) algoritmas (Srikant, Agrawal, 1996) buvo pasiūlytas R. Agrawal ir R. Srikant 1996 m., kuris yra Apriori-algoritmų atstovas ir naudoja kandidatų generavimo, testavimo ir atmetimo strategiją. GSP algoritmas tapo gerai žinomu ir naudojamu algoritmu dažnų sekų paieškoje bei sulaukė daugelio patobulinimų, žymiai pagerinusių algoritmo veikimo greitį.

SPADE (angl. *Sequential Pattern Discovery using equivalence classes algorithm*) algoritmas (Zaki, 2001) buvo pasiūlytas M. J. Zaki 2001 m., kuris naudoja vertikalų dažnų sekų saugojimo formatą taip sumažindamas poreikį daug kartų skaityti pradinę duomenų bazę palyginus su GSP algoritmu. Abu GSP ir SPADE algoritmai naudoja Apriori principą ir generuoja didelį kandidatų sąrašą bei kelis kartus skaito pradinę duomenų bazę.

PrefixSpan algoritmas (Pei et al., 2001) buvo pasiūlytas J. Pei et al. 2001 m., kuris veikia „skaldyk ir valdyk“ principu ir identifikuoja priešdėlines sekas (angl. *prefix*) bei visas sekas duomenų bazėje suskirsto pagal identifikuotas priešdėlines sekas ir rekursiškai apskaičiuoja sekų dažnius. PrefixSpan algoritmas veikia greičiausiai palyginus su SPADE ir GSP.

SPAM (angl. *Sequential Pattern Mining*) algoritmas (Ayres et al., 2002) buvo pasiūlytas J. Ayres et al. 2002 m., kuris naudoja vertikalų bitmat formatą sekoms saugoti, bei taip leidžia efektyviai skaičiuoti sekų dažnius pradinėje duomenų bazėje. PrefixSpan algoritmas veikia greičiau nei SPAM ant mažų duomenų bazių, tačiau didelėse duomenų bazėse SPAM gerokai lenkia PrefixSpan ir SPADE algoritmus. Pagrindinis SPAM algoritmo trūkumas yra tas, kad visa pradinė duomenų bazė su visomis algoritmo naudojamomis struktūromis turi visiškai tilpti į pagrindinę atmintį, todėl algoritmas nėra tinkamas labai didelių duomenų bazių analizei arba nenutrūkstamiems duomenų srautams.

LAPIN (angl. *Last Position Induction algorithm*) algoritmas (Yang et al., 2006) buvo pasiūlytas Z. Yang et al. 2006 m., kuris pasiūlė efektyvų algoritmą kaip sprendimą tankiose (angl. *dense*) duomenų bazėse.

PRISM algoritmas (angl. *Prime-Encoding Based Sequence Mining*) (Guoda et al., 2007; Gouda et al., 2010) buvo pasiūlytas K. Gouda et al. 2007 m., kuris naudoja vertikalią blokinę sekų saugojimo ir dažnių skaičiavimo strategiją, besiremiančią pirminių skaičių faktorizacijos teorija. PRISM algoritmas gerokai lenkia anksčiau pasiūlytus algortimus, tokius kaip SPADE, PrefixSpan ir SPAM. Laiko palyginimo tarp PRISM ir LAPIN algoritmų nebuvo atlikta.

3. Pasiūlyti apytiksliai dažnų sekų paieškos metodai

3.1. Egzistuojančių apytikslių algoritmų apžvalga

Nuo 1995 m. buvo pasiūlyta daug tiksliųjų dažnų sekų paieškos algoritmų bei jų patobulinimų - GSP, SPADE, SPAM, PrefixSpan, LAPIN, PRISM, ir t. t. Visi anksčiau išvardinti algoritmai yra tikslieji dažnų sekų paieškos algoritmai, kurie reikalauja bent kelis kartus nuodugniai nuskaityti pradinę duomenų bazę.

Turint omenyje sparčiai augančius duomenų kiekius tampa akivaizdu, kad daugelyje sričių, kur yra svarbu algoritmo greitis, tik apytiksliai dažnų sekų paieškos algoritmai gali būti pritaikyti duomenims analizuoti. Mūsų žiniai anksčiau pasiūlyti apytiksliai dažnų sekų paieškos algoritmai neturi teorinio paklaidų įvertinimo, o remiasi tik empiriniais algoritmų bandymais skirtingose duomenų bazėse.

ApproxMAP algoritmas (angl. *Approximate Multiple Alignment Pattern mining*) (Kum et al., 2003) buvo pasiūlytas H. C. Kum et al. 2003 m. ApproxMAP algoritmo idėja yra ta, kad užuot ieškojęs tikslių dažnų sekų pradinėje duomenų bazėje, algoritmas identifikuoja sekas, dažnai naudojamas daugelyje kitų sekų. ApproxMAP algoritmo autoriai nepateikia algoritmo greičio tyrimų, tik tikslumo tyrimus, kurie buvo atlikti naudojant tam tikro tipo duomenų bazę. Rezultatai parodė, kad ApproxMAP algoritmas labai tinka rasti įžvalgas pradinėje duomenų bazėje, tačiau jis kelis kartus skaito pradinę duomenų bazę ir labiau susitelkia ties apytikslių sekų radimu nei algoritmo greičiu. Taip pat algoritmas reikalauja gerų žinių apie pradinę duomenų bazę tam, kad būtų gauti naudingi rezultatai.

Apytikslis ProMFS algoritmas (angl. *Probabilistic algorithm for Mining Frequent Sequences*) (Tumasonis, Dzemyda, 2004) buvo pasiūlytas R. Tumasonio ir G. Dzemydos 2004 m., kuris remiasi statistinėmis charakteristikomis pagal tai, kokia tvarka elementai pasirodo pradinėje duomenų bazėje. Šių charakteristikų radimas reikalauja visiško pradinės duomenų bazės nuskaitymo. Algoritmas formuoja daug trumpesnę pradinės duomenų bazės imtį bei ją analizuoja tiksliuoju GSP algoritmu. ProMFS algoritmas remiasi tik empiriniais bandymais ir stebėjimais, kaip algoritmas veikia skirtingose duomenų bazėse, tačiau neturi jokio teorinio daromų paklaidų įverčio.

3.2. Naujas apytikslis atsitiktinės imties metodas (angl. *Random Sampling Method* (RSM))

Atsitiktinės imties metodas (RSM) yra daug spartesnis nei tikslieji dažnų sekų paieškos metodai, nes analizuoja ne visą pradinę duomenų bazę, o daug trumpesnę jos atsitiktinę imtį. RSM metodas yra apytikslis, tačiau jo paklaidų tikimybes galima įvertinti naudojant standartinius statistinius metodus.

Pradinės duomenų bazės atsitiktinė imtis $\bar{S}_n$ 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 sekų, 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 $a_i a_j$ yra $(S_{\eta_1}, S_{\eta_1+1})$, $(S_{\eta_2}, S_{\eta_2+1})$, …, $(S_{\eta_n}, S_{\eta_n+1})$. *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 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 $\eta_1, \eta_2, \dots, \eta_n$.

Pasinaudoję bet kuriuo tiksliuoju dažnų sekų paieškos algoritmu, nustatome sekų $a_{i_1}, a_{i_2}, \dots, a_{i_k}$ empirinius dažnius atsitiktinėje imtyje $\bar{S}_n$.

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

Šioje disertacijoje naudojamas tikslusis GSP algoritmas, tačiau norint dar labiau pagerinti RSM algoritmo greitį būtų galima naudoti PRISM arba LAPIN algoritmus, kurie yra greičiausi tarp tiksliųjų algoritmų. Pasirenkame klasifikavimo paklaidos slenkstį $\delta > 0$ $(0 < \varepsilon - \delta < \varepsilon + \delta < 1), k = 1,2, \dots$. Visas sekas $a_{i_1}, a_{i_2}, \dots, a_{i_k}$ klasifikuojame į tris klases:

1) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \geq \varepsilon + \delta$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame dažnų sekų klasei;
2) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \leq \varepsilon - \delta$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame retų sekų klasei;
3) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \in (\varepsilon - \delta, \varepsilon + \delta)$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame tarpinių sekų klasei.

Aptarsime tikimybinio algoritmo klaidų tikimybių įvertinius. Fiksuokime kokią nors seką $a_{i_1}, \dots, a_{i_k}$. Galimos dviejų rūšių klaidos:

1) seka priskirta dažnų sekų klasei, tačiau iš tikrųjų ji yra reta;
2) seka priskirta retų sekų klasei, tačiau iš tikrųjų ji yra dažna.

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

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

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

$$EZ_i = p,$$

(3)

o dispersija

$$DZ_i = p(1-p).$$

(4)

Tikimybes (1) ir (2) galima įvertinti standartiniais matematinės statistikos metodais: remiantis binominio skirstinio savybėmis grąžintinės imties atveju bei hipergeometrinio skirstinio savybėmis negrąžintinės imties atveju. Apsiribosime asimptotiniais klaidų tikimybių įvertiniais grąžintinės imties atveju. Jie veiksmingi, kai imties didumas n yra pakankamai didelis. Apibrėžkime atsitiktinį dydį

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

Remiantis centrine ribine teorema su visais $a \le b$

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

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 $a \le b$

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

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

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

(5)

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

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

(6)

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

$$\max_{p<\varepsilon} P(\overline{p_n} - p > \delta) \approx \max_{p<\varepsilon}\left(1 - \Phi(\delta\frac{\sqrt{n}}{\sqrt{p(1-p)}})\right) \le$$

$1 - \Phi(\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_0(1-\varepsilon_0)}})$, (7)

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

$$\max_{p\ge\varepsilon} P(\overline{p_n} - p < -\delta) \approx \max_{p\ge\varepsilon}\left(\Phi(-\delta\frac{\sqrt{n}}{\sqrt{p(1-p)}})\right) \le \Phi(-\delta\frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}),$$

(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

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

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

Taigi tik perrinkę visą pradinę duomenų bazę galėsime nustatyti, ar seka $a_{i_1}, \dots, a_{i_k}$ yra dažna ar reta arba naudojant daugybinio perskaičiavimo metodą (MRM). Kita vertus, kad ir koks būtų p, jis yra artimas empiriniam dažniui $\bar{p}_n$, kai n yra pakankamai didelis, nes vėlgi 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 suderinamumas, o jų sąryšį galima išreikšti lygybe $\delta\sqrt{n} = \text{const}$.

3.4. Naujas apytikslis daugybinio perskaičiavimo metodas (angl. *Multiple Re-sampling Method* (MRM))

RSM atsitiktinės imties metodas visas sekas klasifikuoja į tris klases:

1) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \geq \varepsilon + \delta$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame dažnų sekų klasei;
2) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \leq \varepsilon - \delta$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame retų sekų klasei;
3) jeigu $\bar{p}_n(a_{i_1}, \dots, a_{i_k}) \in (\varepsilon - \delta, \varepsilon + \delta)$, tai seką $a_{i_1}, \dots, a_{i_k}$ priskiriame tarpinių sekų klasei.

Tarpinių sekų klasė RSM metode buvo sukurta, nes klasifikavimo paklaida gali būti gana didelė, netgi arti $\frac{1}{2}$. Klasifikavimo paklaidos tikimybė priklauso nuo to, kiek tikrasis sekos dažnis p skiriasi nuo ε, nes pagal centrinę ribinę teoremą, jei $p = \varepsilon$, tuomet neteisingo klasifikavimo tikimybė artėja į $\frac{1}{2}$:

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

Kad nustatytume, ar sekos, priklausančios tarpinių sekų klasei, yra dažnos ar retos, galima visiškai nuskaičius pradinę duomenų bazę ir apskaičiavus tikruosius sekų dažnius, tačiau daugeliu atvejų visiškas duomenų nuskaitymas gali būti labai sudėtingas, jeigu duomenų bazė yra didelė.

Daugybinio perskaičiavimo metodas (MRM) yra atsitiktinės imties metodo (RSM) patobulinimas, kuris naudoja daugybinio perskaičiavimo strategiją ir analizuoja $h > 1$ atsitiktinių imčių iš pradinės duomenų bazės bei nustato, kiek kartų kiekviena seka buvo dažna, reta ir tarpinė:

1. h_1 – kiek kartų seka buvo identifikuota kaip dažna;
2. h_2 – kiek kartų seka buvo identifikuota kaip reta;

3. h_3 – kiek kartų seka buvo identifikuota kaip tarpinė.

Daugybinio perskaičiavimo metodas (MRM) nustato galutinį sekos priskyrimą prie dažnų arba retų sekų klasės, remiantis max (h_1, h_2, h_3):

1. Seka yra dažna, jeigu $h_1 > h_2$ ir $h_1 > h_3$;
2. Seka yra reta, jeigu $h_2 > h_1$ ir $h_2 > h_3$.

Jeigu nė viena iš dviejų anksčiau išvardytų sąlygų nėra tenkinamos ir $max\ (h_1, h_2, h_3) = h_3$, tuomet h_3 - skaičius, nusakantis, kiek kartų seka buvo identifikuota kaip tarpinė, yra eliminuojamas ir seka yra laikoma dažna, jeigu $h_1 > h_2$; priešingu atveju, jeigu $h_2 > h_1$, tuomet seka yra laikoma reta. Jeigu $h_1 = h_2$, tuomet MRM metodas tokią seką laiko dažna, kad nebūtų prarasti potencialūs kandidatai ilgesnių sekų kandidatų generavime. Naudojant daugybinio perskaičiavimo metodą, galutinis sprendimas, ar tarpinė seka yra dažna ar reta, gali būti priimtas be visiško pradinės duomenų bazės nuskaitymo. MRM metodas yra tikslesnis nei RSM metodas, tačiau jis yra h kartų lėtesnis nei RSM metodas.

3.5. Naujas apytikslis Markovo savybe besiremiantis metodas (angl. *Markov Property Based Method* (MPBM))

Markovo savybe besiremiantis metodas kelis kartus skaito pradinę duomenų bazę ir apytiksliai suranda ilgesnių sekų dažnius naudodamas tiksliai identifikuotus tikruosius dažnius. Kad būtų galima nustatyti dažnas sekas pradinėje duomenų bazėje naudojantis pirmosios eilės Markovo procesu, pakanka du kartus nuskaityti pradinę duomenų bazę (antrosios eilės Markovo procese pakanka tris kartus nuskaityti pradinę duomenų bazę ir t. t.).

Laikysime, kad pradinė duomenų bazė yra stacionaraus erdoginio proceso realizacija. Laikysime, jog $X_n, -\infty < n < \infty$ yra stacionarus stochastinis procesas: kiekvienam $n, m, t_1, \dots, t_n$ ir $\alpha_{i_1}, \dots, \alpha_{i_n} \in V$,

$$P(X_{t_1} = \alpha_{i_1}, \dots, X_{t_n} = \alpha_{i_n}) = P(X_{t_1+m} = \alpha_{i_1}, \dots, X_{t_n+m} = \alpha_{i_n})$$

Dėl ergodiškumo savybės (Varadhan, 2001) $\bar{p}(\alpha_{i_1}, \dots, \alpha_{i_m})$ turi ribą ir dėl ergodiškumo savybės riba yra:

$p(\alpha_{i_1}, \dots, \alpha_{i_m}) = P\{X_0 = \alpha_{i_1}, X_1 = \alpha_{i_2}, \dots, X_{m-1} = \alpha_{i_m}\}$:

$$\bar{p}(\alpha_{i_1}, \dots, \alpha_{i_m}) \to p(\alpha_{i_1}, \dots, \alpha_{i_m}), \text{ kai } N \to \infty. \qquad (7)$$

Laikykime, kad $X_n, n = 1,2, \dots$ yra stacionarus pirmosios eilės Markovo procesas. Pažymėkime:

$$p(\alpha_{i_1}\alpha_{i_2} \dots \alpha_{i_k}) = P\{X_n = \alpha_{i_1}, X_{n+1} = \alpha_{i_2}, \dots, X_{n+k-1} = \alpha_{i_k}\}, k = 1,2, \dots$$

Tarkime, kad mes žinome $p(\alpha_i), p(\alpha_i\alpha_j),\ i,j = 1, \dots, M$ kelis kartus nuskaitę pradinę duomenų bazę.

LEMA 1. Laikysime, kad $X_n, n = 1,2, \dots$ yra stacionarus ergodiškas pirmosios eilės Markovo procesas. Tuomet aproksimuojant pradinę duomenų bazę stacionariu pirmosios eilės Markovo procesu (m=1) kiekvienam $\alpha_{i_1}, \dots, \alpha_{i_m} \in V$, mes gauname

$$p(\alpha_{i_1}, \dots, \alpha_{i_m}) = \frac{p(\alpha_{i_1,}\alpha_{i_2})p(\alpha_{i_2},\alpha_{i_3}) \dots p(\alpha_{i_{m-1}},\alpha_{i_m})}{p(\alpha_{i_2}) \dots p(\alpha_{i_{m-1}})}, m \geq 3.$$

Įrodymas. Naudosime Markovo savybę šiai lemai įrodyti. Pasak Markovo savybės, kiekvienam $n \in \{1, \dots, N\}$, k yra Markovo proceso eilė ir $\alpha_{i_1}, \dots, \alpha_{i_n} \in V$

$$P\{X_n = \alpha_{i_n} | X_{n-1} = \alpha_{i_{n-1}}, \dots, X_1 = \alpha_{i_1}\} = P\{X_n = \alpha_{i_n} | X_{n-1} = \alpha_{i_{n-1}}, \dots, X_{n-k} = \alpha_{i_{n-k}}\}$$

Pagal sąlyginės tikimybės apibrėžimą:

$$p(\alpha_{i_1}, \alpha_{i_2}, \alpha_{i_3}) = P\{X_1 = \alpha_{i_1}, X_2 = \alpha_{i_2}, X_3 = \alpha_{i_3}\} =$$

$$= P\{X_3 = \alpha_{i_3} | X_2 = \alpha_{i_2}, X_1 = \alpha_{i_1}\} P\{X_2 = \alpha_{i_2}, X_1 = \alpha_{i_1}\} =$$

$$= P\{X_3 = \alpha_{i_3} | X_2 = \alpha_{i_2}\} P\{X_2 = \alpha_{i_2}, X_1 = \alpha_{i_1}\} =$$

$$= \frac{P\{X_3 = \alpha_{i_3}, X_2 = \alpha_{i_2}\}}{P\{X_2 = \alpha_{i_2}\}} P\{X_2 = \alpha_{i_2}, X_1 = \alpha_{i_1}\} = \frac{p(\alpha_{i_1,}\alpha_{i_2})p(\alpha_{i_2}, \alpha_{i_3})}{p(\alpha_{i_2})}$$

Kad būtų paprasčiau, analizuojamas tik $m = 3$ atvejis, tačiau kai $m > 3$, įrodymas yra analogiškas. Taigi aproksimuojant duomenų bazę stacionariu pirmosios eilės Markovo procesu mes turime žinoti tik šiuos tikruosius dažnius $p(\alpha_{i_1})$ ir $p(\alpha_{i_1}, \alpha_{i_2})$ visiems $\alpha_{i_1}, \alpha_{i_2} \in V$. Lema įrodyta.

Analogiškas prielaidas galime daryti ir naudojant aukštesnės eilės Markovo procesą, tačiau tuomet atitinkamai daugiau kartų reikės skaityti pradinę duomenų bazę, kad būtų galima identifikuoti tikruosius dažnius (pvz. antrosios eilės Markovo procese reikia skaityti pradinę duomenų bazė tris kartus).

4. Pasiūlytų apytikslių dažnų sekų paieškos metodų eksperimentiniai tyrimai

Eksperimentiniai tyrimai disertacijoje pasiūlytiems RSM, MRM ir MPBM metodams buvo atliekami naudojant tikrus ir dirbtinai sugeneruotus duomenis bei asmeninį kompiuterį su 2.4 GHz Intel Celeron procesoriumi, 4GB RAM atminties. Visi metodai realizuoti naudojant Java programavimo kalbą.

4.1. Eksperimentinės duomenų bazės

Eksperimentiniai tyrimai buvo atliekami naudojant tikrą finansinę duomenų bazę bei dirbtinai sugeneruota duomenų bazę:

1. Finansinė duomenų bazė, kuri susideda iš EUR-USD valiutų poros kurso valandinių duomenų nuo 03/01/2000 iki 01/05/2013 (duomenys paimti iš *Online Trading Platform MetaTrader 4 History Center* (MetaTrader 4 software, retrieved 2013)). Finansinė duomenų bazė susideda iš $N = 5397843$ elementų, kurie įgyja reikšmes {A, B, C}, kurios nurodo, ar valiutų kursas auga, krenta ar lieka nepakitęs palyginus su praėjusia valanda:

- 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 $C_i = O_i$.

2. Genetinė duomenų bazė susideda iš $N = 66326866$ elementų, kurie įgyja galimas reikšmes {A, T, C, G}, kurios reiškia DNA nukleobazes (cytosine, thymine, adenine, guanine). Duomenys buvo sugeneruoti naudojantis DNA Baser Sequence Assembler software (DNA Baser, retrieved in 2013).

4.2. Parametrų parinkimas disertacijoje pasiūlytiems apytiksliams metodams

Pradinės duomenų bazės S analizei naudosime GSP algoritmą bei pasiūlytus RSM, MRM ir MPBM metodus dažnų sekų paieškai pradinėje duomenų bazėje. Žemiau išvardyti vartotojo nurodyti parametrai bus naudojami anksčiau išvardytuose metoduose:

- Dažnio slenkstis (angl. *minimum support threshold*) $0 < \varepsilon < 1$, kuris naudojamas nustatyti, ar seka yra dažna, ar reta. Seka yra laikoma dažna, jeigu $\bar{p}_n \geq \varepsilon$; priešingu atveju ji laikoma reta. Dažnio slenkstis yra naudojamas visuose dažnų sekų paieškos metoduose: GSP, RSM, MRM, MPBM, ProMFS;
- Atsitiktinės imties, sudarytos iš pradinės duomenų bazės, dydis n, kuris yra naudojamas RSM ir MPBM metoduose;
- Klasifikavimo klaidos slenkstis $\delta > 0$, toks, kad $(0 < \varepsilon - \delta < \varepsilon + \delta < 1)$. Slenkstis naudojamas apytiksliuose RSM ir MRM metoduose;
- Atsitiktinės imties, sudarytos iš pradinės duomenų bazės, perskaičiavimų skaičius $h = 30$. Šis parametras naudojamas MRM metode;
- Markovo proceso eilė $m = 1, 2, 3$, naudojama apytiksliame MPBM metode.

4.3. RSM, MRM ir MPBM metodo eksperimentinis tyrimas

Tikslusis GSP algoritmas ir aspytikslis RSM, MRM ir MPBM metodai bus naudojami dažnų sekų paieškai pradinėje duomenų bazėje S. Laikysime, kad seka yra dažna, jeigu jos dažnis didesnis arba lygus $\varepsilon = 0{,}08$. Pirmiausia, nagrinėsime pradinę duomenų bazę naudojant GSP algoritmą bei tiksliai nustatysime sekų tikruosius dažnius. Tuomet naudojant apytikslius metodus nustatysime empirinius sekų dažnius analizuodami atsitiktinę pradinės duomenų bazės imtį, kurios dydis yra $n = 500$, $n = 1000$ ir $n = 2000$, o klasifikavimo klaidos slenkstis yra $\delta = 0{,}02$.

RSM metodo pirmojo tipo klasifikavimo klaidos tikimybė (seka yra dažna, tačiau buvo priskirta retų sekų klasei):

$n = 500$: $\quad 1 - \Phi\left(\delta \frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(1.6485) \approx 0.0496;$

$n = 1000$: $\quad 1 - \Phi\left(\delta \frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(2.3313) \approx 0.0099;$

$n = 2000$: $\quad 1 - \Phi\left(\delta \frac{\sqrt{n}}{\sqrt{\varepsilon(1-\varepsilon)}}\right) = 1 - \Phi(3.2969) \approx 0.0005.$

RSM metodo antrojo tipo klasifikavimo klaidos tikimybė (seka yra reta, tačiau buvo priskirta dažnų sekų klasei):

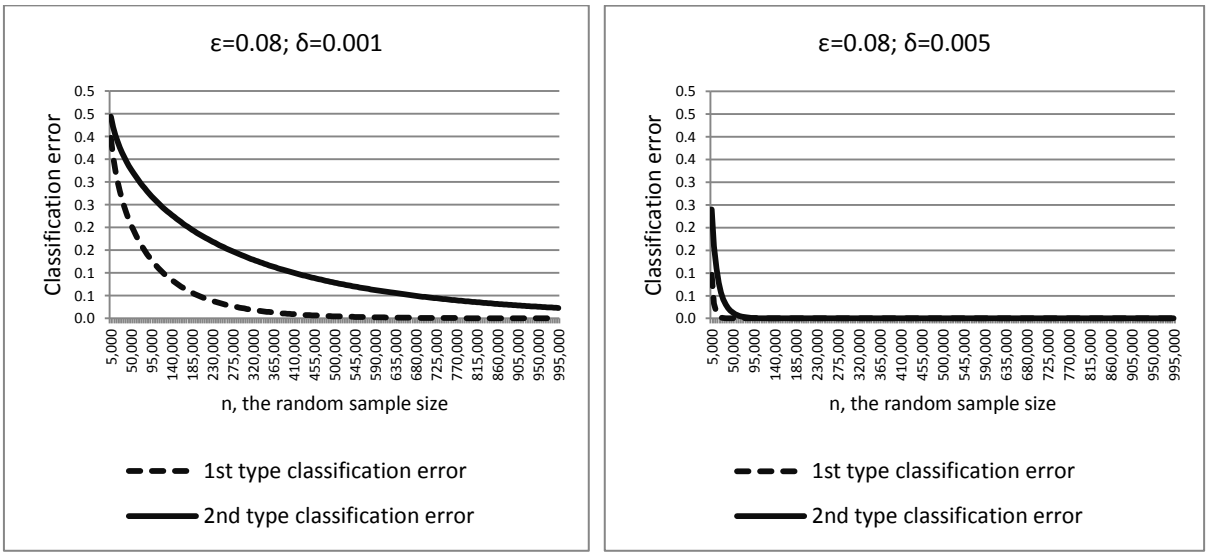
$n = 500$: $\quad \Phi\left(-\delta \frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-0.8944) \approx 0.1855;$

$n = 1000$: $\quad \Phi\left(-\delta \frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-1.2649) \approx 0.1030;$

$n = 2000$: $\quad \Phi\left(-\delta \frac{\sqrt{n}}{\sqrt{\varepsilon_1(1-\varepsilon_1)}}\right) = \Phi\left(-2\delta\sqrt{n}\right) = \Phi(-1.7889) \approx 0.0368.$

Svarbu pastebėti, jog augant atsitiktinės imties dydžiui n auga ir tarpinių sekų skaičius. Tarpinių sekų skaičius gali būti sumažintas mažinant klaisfikavimo klaidos slenkstį δ, bet tuomet pirmojo ir antrojo tipo klaidų tikimybės auga. Jos gali būti sumažintos didinant n. Taigi suderinamumas tarp δ ir n yra būtinas, o jų sąryšis apibrėžiamas kaip

$\delta\sqrt{n} = \text{const}$. 1 pav. pavaizduota priklausomybė tarp n ir $\delta = 0.001$ ir 0.005, o dažnių slenkstis $\varepsilon = 0.08$.

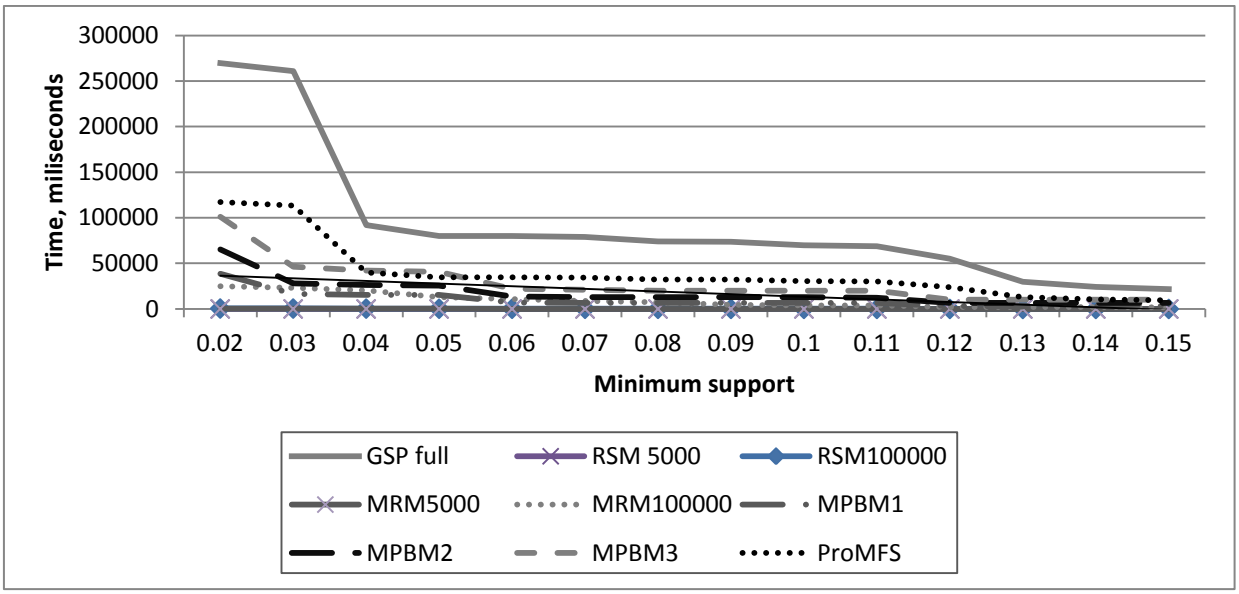


1 pav. Pirmojo ir antrojo tipo klasifikavimo klaidos, kai $\varepsilon = 0.08; \delta = 0.001$ ir 0.005.

MPBM metode naudosime m = 1, 2 ir 3 eilės Markovo procesus, o MRM metode sudarysime $h = 30$ atsitiktinių pradinės duomenų bazės imčių analizei.

Lentelė 1. MPBM (m = 1, 2 and 3), RSM ($n = 1000$ and $n = 2000$; $\delta = 0{,}02$) ir MRM ($h = 30$) metodų rezultatai

Duomenų bazė	Rezultatai	GSP algoritmas	MPBM metodas			RSM metodas		MRM metodas	
			m=1	*m=2*	*m=3*	*n*=1000	*n*=2000	*n*=1000	*n*=2000
Finansinė duomenų bazė	Metodo surastų dažnų sekų skaičius	265	268	266	266	193	212	261	263
	Retų sekų, kurios buvo identifikuotos kaip dažnos, skaičius	0	6	4	2	25	17	30	22
	Dažnų sekų, kurios buvo identifikuotos kaip retos, skaičius	0	3	3	1	9	5	13	7
	Tarpinių sekų skaičius	0	0	0	0	163	86	0	0
Genetine duomenų bazė	Metodo surastų dažnų sekų skaičius	144	144	146	146	126	149	139	146
	Retų sekų, kurios buvo identifikuotos kaip dažnos, skaičius	0	0	20	8	12	4	15	6
	Dažnų sekų, kurios buvo identifikuotos kaip retos, skaičius	0	0	18	6	9	6	13	8
	Tarpinių sekų skaičius	0	0	0	0	126	35	0	0



2 pav. Apytikslių metodų greičio palyginimas, kai $\delta = 0.1$.

Eksperimentų rezultatai parodė, kad MRM metodas pademonstravo tiksliausius rezultatus abiejose duomenų bazėse; antroje vietoje pagal tikslumą liko MPBM metodas naudojantis trečiosios eilės Markovo procesu; tačiau šis metodas reikalauja skaityti pradinę duomenų bazę keturis kartus, todėl jis gana sunkiai gali būti naudojamas labai didelėse duomenų bazėse arba nenutrūkstamuose duomenų srautuose. RSM metodas neskaito pradinės duomenų bazės ir yra greičiausias iš analizuotų metodų, tačiau jis turi gana didelį skaičių tarpinių sekų ir tikslumu jį lenkia ir MPBM, ir MRM metodai. MRM metodas eliminavo tarpines sekas ir pademonstravo tiksliausius rezultatus, tačiau jis dirba $h = 30$ kartų ilgiau palyginus su RSM metodu, nes analizuoja 30 atsitiktinių imčių iš pradinės duomenų bazės.

Didelis RSM metodo privalumas yra teorinis paklaidų įvertinimas naudojant standartinius statistinius metodus ir centrinę ribinę teoremą. Teorinis paklaidų įvertis leidžia žinoti metodo daromas paklaidas iš anksto bei nereikalauja išplėstinių empirinių bandymų metodo tikslumui nustatyti. RSM metodas neskaito pradinės duomenų bazės taip grąžindamas greitus rezultatus bei yra tinkamas naudoti daugelyje sričių, kuriose metodo daromos paklaidos yra priimtinos.

5. Interneto vartotojų elgsenos duomenų analizė ir vizualizavimas

5.1. Interneto vartotojų elgsenos duomenys

Interneto vartotojų elgenos analizė yra svarbus uždavinys, leidžiantis padidinti reklaminių kampanijų internete pelną bei efektyvumą. Reklaminių kampanijų duomenų analizės rezultatai yra peržiūrimi žmonių, neturinčių techninio išsilavinimo, todėl jie turi būti pateikti pačių patogiausiu ir suprantamiausiu būdu. Interneto vartotojų skaičius išaugo eksponentiškai, o kartu pasikeitė ir vartotojų įpročiai – vis daugiau laiko praleidžiama internete bei vis daugiau prekių perkama internetu. Pasak IAB apžvalgos, europiečiai vidutiniškai praleidža 24 valandas internete kiekvieną mėnesį (IAB digital review: Consumers driving the digital uptake, 2010).

Visi pirkėjai internetu gali būti klasifikuojami į apsilankančius pardavėjo puslapyje tiesiogiai arba per reklaminės kampanijos medžiagą. Įvairi reklaminė medžiaga internete gerokai padidina pirkėjų skaičių pardavėjo puslapyje. Pasak comScore apžvalgos (comScore: The 2010 Europe Digital Year in Review, 2011) 97% visų interneto vartotojų Europoje matė reklaminę medžiagą naršydami internete. Kiekviena reklaminė kampanija surenka didžiulį kiekį duomenų apie interneto vartotojus, mačiusius reklaminę medžiagą, pvz., kiek kartų vartotojas matė arba paspaudė ant reklaminės medžiagos, kiek produktų rinkosi pardavėjo puslapyje, kokius produktus nusipirko, kiek už juos mokėjo ir t. t. Visi šie duomenys yra sekami naudojantis slapukais (angl. *cookies*) ir pagal statistiką 87% interneto vartotojų gali būti sekami naudojant naršyklės slapukus (Visitor Browser & Cookie Demographics, retrieved 2013).

Šiame skyriuje yra analizuojami duomenys iš dviejų realių reklaminių kampanijų, kurios reklamavo vieno pardavėjo prekes (5361 pardavimų) naudojant įvairią reklaminę medžiagą, kuri buvo rodyta populiariausiuose tinklalapiuose bei paieškos sistemose (Google, Yahoo, MSN). Pardavėjo puslapis buvo nuodugniai sekamas įskaitant įsigytų prekių kainas. Kiekvienas pirkimas pardavėjo puslapyje buvo nusakomas 9 skirtingais požymiais $x_1,\ x_2 \ldots, x_9$.

Penki kiekybiniai požymiai:

- x_1 – kampanijos tipas: 1 – visi vartotojai, kurie apsilankė puslapyje, pamatę reklaminę medžiagą (angl. *campaign traffic*), 0 – visi vartotojai, kurie apsilankė pardavėjo puslapyje tiesiogiai ir nematė reklaminės medžiagos (angl. *non-campaign traffic*);
- x_2 – reklaminės medžiagos dydis cm^2;
- x_3 – minutės nuo paskutinės sąveikos su reklamine medžiaga;
- x_4 – skaičius, kiek kartų vartotojas matė, spaudė ar kitaip sąveikavo su reklamine medžiaga;
- x_5 – pirkinių kaina.

Keturi kokybiniai požymiai:

- x_6 – reklaminės kampanijos pavadinimas;
- x_7 – sąveikos su reklamine medžiaga tipas;
- x_8 – nukreipiantis tinklalapis, iš kurio vartotojas atėjo į pardavėjo puslapį;
- x_9 – reklaminės medžiagos tipas (tiesioginiai vartotojai, kurie nematė reklaminės medžiagos; interaktyvi reklaminė medžiaga (angl. *rich media*), paieškos žodžiai (angl. *search keywords*), ir t. t.).

Visi kiekybiniai požymiai buvo normuoti, kad jų reikšmės atitiktų intervalą [0, 100].

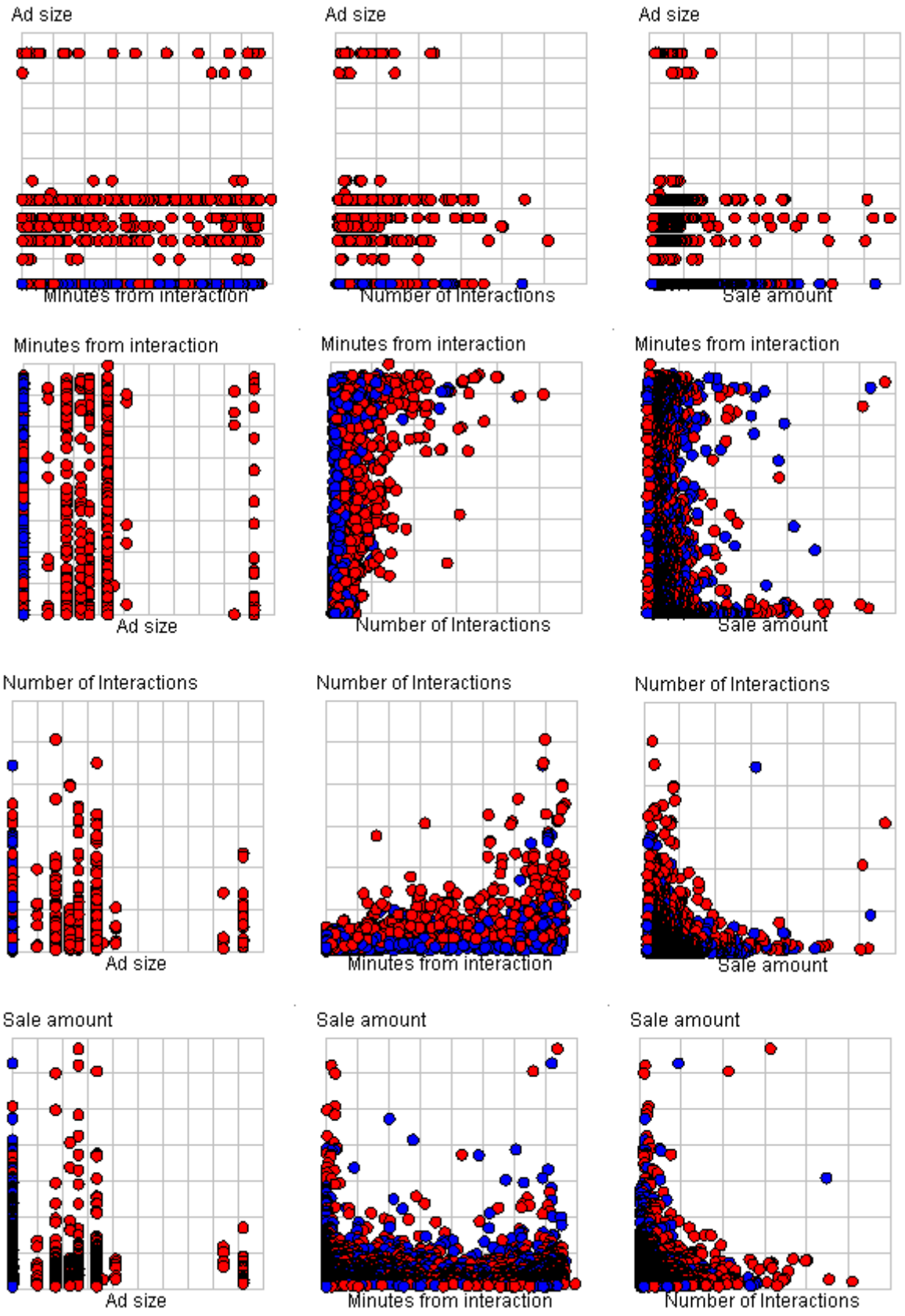
5.2. Interneto vartotojų elgsenos vizualizavimo metodai

Interneto vartotojų elgsenos duomenys buvo analizuojami ir vizualizuoti žemiau išvardintais būdais:

- Geometriniais metodais (taškiniais grafikais, angl. *scatter plots*);
- Daugiamačių skalių algoritmu (Sammono algoritmu);
- Neuroniniais tinklais (saviorganizuojančiu neuroniniu tinklu).

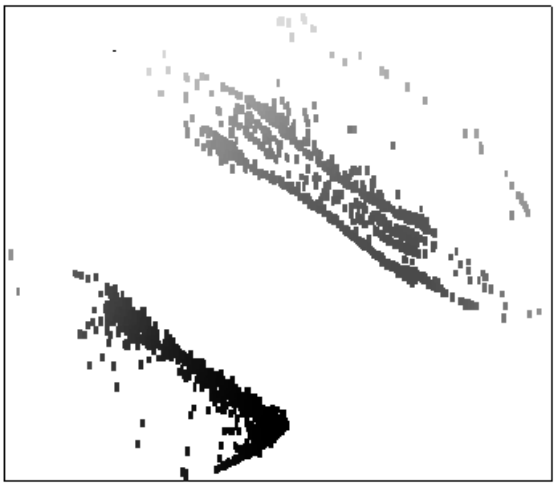
Taškiniai grafikai vizualizuoja priklausomybes tarp pasirinktų požymių bei leidžia pastebėti išsiskiriančius taškus, klasterius arba tendencijas duomenyse. Taškinių grafikų matrica leidžia apžvelgti visas 2-dimensijų koreliacijas tarp skirtingų požymių. 3 pav. pavaizduota kiekybinių požymių taškinių grafikų matrica, kurioje spalva reiškia reklaminės kampanijos tipą (raudona – vartotojai, matę reklaminę medžiagą; mėlyna – vartotojai, nematę reklaminės medžiagos), naudojant RapidMiner programinę įrangą (Rapid Miner software, retrieved 2013). Taškinių grafikų matricos privalumas yra tas, kad poromis vizualizuoti požymiai gali būti lengvai interpretuojami, tačiau augant dimensijų skaičiui, lieka mažai vietos kiekvienai projekcijai bei didelis duomenų kiekis gali būti sunkiai suprantamas iš vizualizacijos. Remiantis 3 pav. vizualizacija, gali būti padarytos tokios išvados:

- Dauguma pirkėjų turėjo tik kelias sąveikas (matė arba paspaudė) su reklamine medžiaga prieš perkant pardavėjo puslapyje.
- Dauguma pirkėjų pirko pardavėjo puslapyje po gana trumpo laiko tarpo nuo to, kai jie paskutinį kartą matė arba paspaudė ant reklaminės medžiagos.
- Vidutinio dydžio reklaminė medžiaga buvo dažniausiai matyta tarp visų reklaminės medžiagos tipų
- Brangiausiai perkantys pirkėjai paprastai labai greitai nusipirko prekes po reklaminės medžiagos pamatymo arba apsilankė pardavėjo puslapyje nematę reklaminės medžiagos iš viso.
- Nėra aiškių pirkėjų klasterių, kuriuos būtų galima įžvelgti naudojant taškinių grafikų matricą.



3 pav. Kiekybinių požymių vizualizacija naudojant taškinių grafikų matricą.

Interneto vartotojų elgsenos vizualizavimui daugiamačių skalių algoritmu (Sammono algoritmu) buvo naudojami tik kiekybiniai požymiai, kurie buvo pavaizduoti dvimatėje erdvėje ($d = 2$). Interneto vartotojų vizualizacija yra pateikta 4 pav., naudojant Orange 2.0 programinę įrangą (Orange 2.0 software, retrieved in 2013). Vizualizacijos rezultatuose galima aiškiai įžvelgti skirtingus klasterius, kurie atitinka skirtingas pirkėjų grupes, kurios yra glaudžiai susijusios viena su kita, jei yra atvaizduotos arti viena kitos grafike. Apatinis klasteris 4 pav. atitinka visus pirkėjus, kurie apsilankė pardavėjo puslapyje tiesiogiai ir neturėjo jokios sąveikos su reklamine medžiaga. Kiti 5 pirkėjų klasteriai suskirstyti pagal reklaminės medžiagos dydį: viršutinis klasteris atitinka didžiausią reklaminę medžiaga, o žemiausias – mažiausią.



4 pav. Kiekybinių požymių vizualizacija naudojant Sammono algoritmą.

Didelis daugiamačių skalių algoritmo privalumas yra tas, kad jis parodo sąryšius tarp duomenų: daugiamačiai duomenys, kurie yra panašūs tarpusavyje bus pavaizduoti arti vienas kito vizualizacijoje. O daugiamačių skalių algoritmo trūkumas yra tas, kad algoritmas užima daug laiko, kai vizualizuojami dideli duomenų kiekiai.

Interneto vartotojų elgsenos analizei buvo naudojamas saviorganizuojantis neuroninis tinklas (angl. *self organizing map (SOM)*), kuris buvo realizuotas naudojantis Viscovery

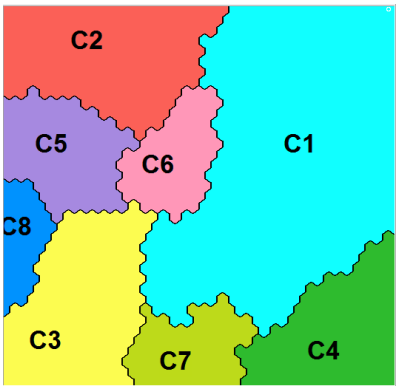
SOMine programine įranga (Viscovery SOMine, retrieved 2013) su šešiakampe tinklo topologija ir Ward kaimynystės funkcija (A ir B yra klasteriai):

$$d(A,B) = \sum_{i \in A \cup B} \|X_i - \vec{m}_{A \cup B}\|^2 - \sum_{i \in A} \|X_i - \vec{m}_A\|^2 - \sum_{i \in B} \|X_i - \vec{m}_B\|^2 = \frac{n_A n_B}{n_A + n_B} \|\vec{m}_A - \vec{m}_B\|^2$$

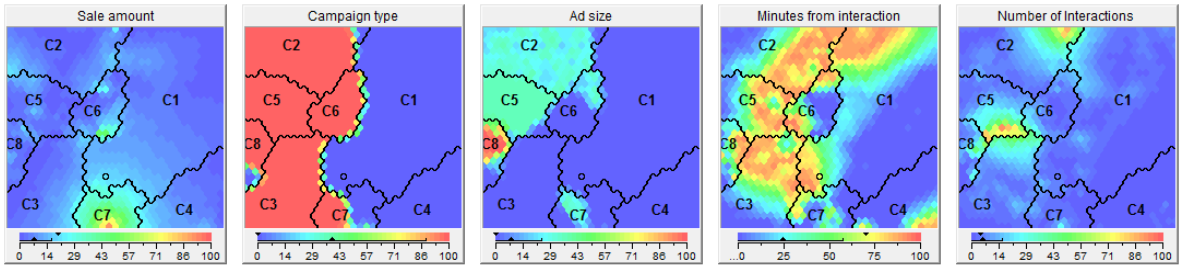
kur $\vec{m}_j$ yra j klasterio centras, o n_j yra taškų skaičius jame. $d(A,B)$ yra vadinama jungimo kaina, kuri atsiranda jungiant du klasterius A ir B.

Saviorganizuojantis neuronis tinklas leidžia lengvai analizuoti ir kokybinius, ir kiekybinius duomenis kartu, nors ir reikalauja žinių pasirenkant papildomus parametrus bei skirtingus svorius, kad būtų gauta kuo geresnė vizualizacija.

Naudojant saviorganizuojantį neuroninį tinklą buvo gauti 8 klasteriai, kurie yra pavaizduoti 5 ir 6 pav. Aukštesnis prioritetas (1.5) buvo duotas pirkimo bendrai sumai, tuo tarpu visi kiti požymiai buvo traktuojami tuo pačiu (1.0) prioritetu. C7 buvo identifikuotas kaip pats pelningiausias pirkėjų klasteris, o jo klientai pirko brangiausias prekes pardavėjo puslapyje bei atliko vidutiniškai 5 sąveikas su reklamine medžiaga prieš pirkdami ir vidutiniškai nuspirko prekes per 20 minučių nuo paskutinės sąveikos su reklamine medžiaga. Beveik pusė C7 klasterio pirkėjų apsilankė puslapyje tiesiogiai, nematę ir nespaudę ant reklaminės medžiagos, o kita pusė dažniausiai ieškojo prekių paieškos sistemose (Google, Yahoo, MSN) ir nematė reklaminės medžiagos.



5 pav. Saviorganizuojančio neuroninio tinklo sukurti klasteriai.



6 pav. Klasteriai, atvaizduoti pagal skirtingus kiekybinius požymius.

Bendrosios išvados

Šios disertacijos tyrimų objektas yra dažnų sekų paieška bei vizualizavimas didelėse duomenų bazėse. Dažnų sekų paieškai didelėse duomenų bazėse buvo pasiūlyti trys nauji apytiksliai (tikimybiniai) metodai: atsitiktinės imties metodas (Random Sampling Method – RSM), daugybinio perskaičiavimo metodas (Multiple Re-sampling Method – MRM) and Markovo savybe besiremiantis metodas (Markov Property Based Method – MPBM). Jų eksperimentinis tyrimas buvo atliktas naudojant tikrą bei dirbtinai sugeneruotą duomenų bazes bei rezultatai palyginti su kitas egzistuojančiais algoritmais. Interneto vartotojų elgsenos duomenys buvo analizuojami bei vizualizuoti naudojant geometrinius metodus, daugiamačių skalių algoritmą bei neuroninius tinklus. Šioje disertacijoje atlikti tyrimai leido padaryti tokias išvadas:

1) Apytiksliai dažnų sekų paieškos metodai gali daug greičiau nei tikslieji dažnų sekų paieškos metodai nustatyti, kurios sekos yra dažnos ir retos. Tai ypač svarbu sparčiai augant duomenų kiekiui, kurį darosi sudėtinga greitai apdoroti, o daugeliu atvejų beveik neįmanoma visiškai perskaityti.
2) Naujai pasiūlytas atsitiktinės imties metodas (RSM) turi didelį privalumą – teoriškai nustato metodo daromas klaidų tikimybes, kurios yra įrodytos naudojant standartinius statistinius metodus. Teorinis paklaidų įvertis leidžia naudoti metodą be išplėstinių empirinių tyrimų naudojant skirtingas duomenų bazes, kurios yra naudojamos kituose dažnų sekų paieškos algoritmuose.
3) Naujai pasiūlytas daugybinio perskaičiavimo metodas (Multiple Re-sampling Method – MRM) pademonstravo tiksliausius rezultatus iš pasiūlytų apytikslių metodų, tačiau

greičiu jis atsilieka nuo atsitiktinės imties metodo (RSM) tiek kartų, kiek skirtingų atsitiktinių imčių yra papildomai analizuojama palyginus su atsitiktinės imties metodu.

4) Naujai pasiūlytas apytikslis Markovo savybe besiremiantis metodas (Markov Property Based Method – MPBM) kelis kartus skaito pradinę duomenų bazę (priklausomai nuo to, kelintos eilės Markovo procesas yra naudojamas, tiek kartų yra skaitoma pradinė duomenų bazė) ir todėl netenkina greičio reikalavimų analizuojant dideles duomenų bazes.
5) Vizualizuojant interneto vartotojų elgsenos duomenis saviorganizuojantis neuroninis tinklas parodė geriausius rezultatus, tačiau jį naudojant vartotojas turi turėti geras žinias apie naudojamus duomenis bei teisingai pasirinkti saviroganizuojančio tinklo naudojamus parametrus. Remiantis vizualizacijos rezultatais gali būti priimami sprendimai apie tolimesnę reklamos strategiją.

Disertacijos autoriaus publikacijų sąrašas

1. J. Pragarauskaite, G. Dzemyda. Visual Decisions in the Analysis of Customer Online Shopping Behaviour. Nonlinear Analysis: Modelling and Control, Vol. 17, ISSN: 1392-5113, No. 3, p. 355–368, 2012.
2. J. Pragarauskaite, G. Dzemyda. Markov Models in the analysis of frequent patterns in financial data. Informatica, Vol. 24, ISSN: 0868-4952, No. 1, p.87 – 102, 2013.
3. J. Pragarauskaitė, G. Dzemyda. Tikimybinis dažnų posekių paieškos algoritmas. Informacijos mokslai, Nr. 50, ISSN 1392-0561, p. 352-357, 2009.
4. J. Pragarauskaite, G. Dzemyda. Probabilistic algorithm for mining frequent sequences. In Proceedings ASMDA 2011, Sapienza University of Rome, p. 1454-1460. Edizioni ETS, 2011.
5. J. Pragarauskaitė, G. Dzemyda. Tikimybinis dažnų posekių paieškos algoritmas. Lietuvos matematikos rinkinys, 51, p. 313–318, ISSN 0132-2818, 2010.

FREQUENT PATTERN ANALYSIS FOR DECISION MAKING IN BIG DATA

Research Context and Motivation

The ever increasing amounts of digital information stored around the World today has the potential to bring significant benefits to people and make it possible to do many things that could not be done previously: forecast financial trends, diagnose and prevent diseases, identify crime suspects, etc. The amount of digital information increases enormously every year and according to the Moore's law, the processing power and storage capacity of computer chips double approximately every 18 months. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 quintillion bytes of data were created (IBM review on what is big data, retrieved 2013). Keeping in mind the amount of raw data, today's algorithms and powerful computers can reveal new important insights that could previously remain hidden.

Frequent pattern mining is a very important task in data mining, especially in big data that consist of millions of records (Han, Kamber, 2006). The term "big data" is used very often recently in data mining and it defines a collection of data sets that are so large and complex that it becomes difficult to process them using on-hand database management tools or traditional data processing applications. The challenges include the capture of data, storage, search, sharing, transfer, analysis, and visualization. Data analysts regularly encounter limitations due to big data in many areas, including meteorology, genomics (Editorial review in Nature, 2008), complex physics simulations, biological data, environmental data, Internet search, finance and business informatics. Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among big data. Moreover, it helps in data classification, clustering, and other data mining tasks as well. Thus, frequent pattern mining has become an important data mining task and attracted a number of different researches in

frequent pattern mining. Exact frequent sequence mining methods deliver accurate results on frequent and rare sequences, however they typically require multiple passes over the original database and, if the database is large, it becomes a time consuming and expensive task. Approximate frequent sequence mining with an estimated probability of error is acceptable in many applications, e.g. marketing, internet user behaviour, stock market, biological databases, etc., because such algorithms are employed in scenarios where timely analysis result is more important than high precision. For example, if an investor can obtain all the approximate frequent sequences from her stock data quickly, then such approximate results might be sufficient to supplement the investor's own investment strategy and to allow the investor to take an optimal and profitable decision on a specific investment. On the other hand, if a stock investor attempts to extract all the accurate frequent sequences from the stock data, it would be very time-consuming and by the time a decision needs to be made, the best time for the investment might have passed. Combining different trading strategies with the knowledge of frequently occurring sequences might be a key to the trader's success; however this is typically a challenging task because frequent patterns have to be identified as quickly as possible. Firstly, the amount of data is typically so huge that it becomes difficult to both store and process it within an acceptable. Secondly, previously acquired knowledge may age as the time goes by and lose its importance. Therefore, in financial markets the computing speed of finding frequent patterns in big data is very important as is using the approximate frequent sequence mining algorithms with an acceptable error threshold. This approach could help making quick decisions with higher profitability on the trading strategy. Frequent sequences have to be identified and presented to the trader as quickly as possible as every second is important to decide whether to open/exit the trading position and the delay could cause that the right moment of opening the trading position might have passed.

Approximate frequent sequence mining methods are much faster than exact methods because instead of doing multiple passes over the original database they analyze a much shorter sample of the original database or are using specific assumptions on the structure of the original database. In many cases finding an exact result in frequent sequence mining is not compatible with limited availability of resources and real time constraints,

but an approximation of the exact result is enough for most purposes, e.g. biological data analysis, stock market analysis, behaviour analysis in the internet, etc. In the past decade there were a number of approximate frequent sequence mining algorithms proposed (ApproxMAP (Pei et al., 2003), ProMFS (Tumasonis, Dzemyda, 2004), approximate frequent sequence mining in data streams (Silvestri, Orlando, 2007)), that are much faster compared to exact frequent sequence mining algorithms.

All state-of-the-art approximate frequent sequence mining methods do not use a theoretical estimate of the probability of error made by the method when identifying frequent patterns in the original database. These methods provide only the empirical evidence based on extensive experiments and observations of algorithm results on different databases.

In this study we propose novel approximate frequent sequence mining methods with a theoretical estimate of the probability of error made by proposed methods when classifying sequences as frequent or rare. The performance of methods is compared to other approximate and exact frequent sequence mining methods and recommendations given on how to select parameters used in proposed methods to achieve more accurate results.

Tasks and Objectives of the Research

In this study the objectives are:

(1) create novel approximate frequent sequence mining algorithms for which theoretical approximation and estimation of induced errors can be made;

(2) propose an approach to visualise big data.

In pursuit of the above objectives, the specific tasks are:

- Study the existing frequent sequence mining algorithms (both exact and approximate);

- Propose novel approximate frequent sequence mining method with an estimate of the probability of error made by this method when classifying the sequences as frequent or rare;
- Construct of a sample of an original database for proposed approximate method;
- Propose random sampling based methods for mining frequent sequences with the estimated probabilities of errors made by these methods;
- Propose a method that relies upon the Markov property for approximately mining frequent sequences;
- Evaluate the performance of proposed approximate methods and compare the results with other exact and approximate frequent sequence mining approaches;
- Visually represent the analysis of behaviour of Internet user using various visualization methods.

Proposed Solutions and Contributions of Scientific Novelty

As a solution to the previously described problem the following three novel frequent sequence mining methods for dealing with big data and visualization of internet users' behaviour are proposed in this thesis:

- Random Sampling Method (RSM) with estimated classification errors using the central limit theorem;
- Multiple Re-sampling Method (MRM) which is an improved version of RSM method that includes re-sampling strategy;
- Markov Property Based Method (MPBM) for frequent sequence mining;
- Visual analysis of behaviour of Internet users.

All three proposed methods for frequent sequence mining are approximate and deliver approximate results in identifying frequent and rare sequences in the original database. All proposed methods are implemented and tested on real and artificial databases and their performance compared with other approximate and exact methods. Their theoretical

error bounds are formulated and proved. The recommendation which parameters to choose by proposed methods is given to the reader in order to get more precise results using these methods.

For visual representation of big data, the behaviour of Internet users was analyzed using geometric methods, multidimensional scaling and artificial neural networks. In visual representation of big data, SOM visualization showed the best results when visualizing the behaviour of Internet users; however it requires some training and experience by a user, who is analyzing the data. Based on the visualization, the decisions on the advertising strategy could be taken by a non-technical person who is managing the advertising campaigns.

Defended Statements

1) Approximate (probabilistic) frequent sequence mining methods could deliver the acceptable results with right balance between the computing speed and precision when dealing with significantly growing amount of digital information in big data; whereas the exact frequent sequence mining methods work under the time constraints and cannot guarantee the required speed by various real-world applications.
2) Random Sampling Method (RSM) brings a significant benefit among existing approximate frequent sequence mining methods by providing a theoretical estimation of error probabilities. The error probabilities made by this method could be theoretically estimated using standard statistical methods. The theoretical estimation allows using the proposed method without any extensive empirical experiments used in other state-of-the-art methods to evaluate the performance of the method.
3) Random Sampling Method may be improved using a re-sampling strategy and eliminating the class of intermediate sequences as proposed in Multiple Re-sampling Method (MRM). The elimination of the class of the intermediate sequences delivers more precise results compared to Random Sampling Method, however requires more computing time (as much as the number of re-samples analyzed).
4) Approximate frequent sequence mining methods that require reading the entire original database do not satisfy the time constraints required by many real-world applications. Therefore Markov Property Based Method (MPBM) proposed in this

thesis, does not satisfy the time constraints as it requires reading the original database at least several times (the number of times equals to the order of the Markov process) and it delivers similar precision results as Multiple Re-sampling Method (MRM).

5) In the visual representation of Internet users' behavioural data, SOM visualization gives the best results; however it requires some training and experience by a user, who is analyzing the data.

Approbation and Publications of the Research

The main results of the dissertation were published in 5 scientific papers: 4 articles in the periodical scientific publications; 1 article in the proceedings of scientific conference. The main results of the work have been presented and discussed at 5 national and international conferences.

Outline of the Dissertation

In Chapter 1 an introduction to frequent sequence mining is provided, including the definition and related work in frequent sequence mining algorithms. In Chapter 2 detailed information about the most popular exact frequent sequence mining algorithms is presented. In Chapter 3 detailed information about the proposed novel approximate frequent sequence mining methods is presented. In Chapter 4 the performance studies of the proposed methods and the comparisons with other approximate and exact frequent sequence mining methods are presented. In Chapter 5 the methodology for analysis of the behaviour of Internet users is presented and examined.

Conclusions

This thesis focuses on several data mining tasks related to analyzing big data: frequent pattern mining and visual representation of data. For mining frequent patterns in big data, three novel approximate (probabilistic) methods were proposed: RSM (Random Sampling Method), MRM (Multiple Re-sampling Method) and MPBM (Markov Property Based Method). Their performance in terms of the computation speed and precision was evaluated on real and artificial databases and compared with other existing state-of-the-art frequent pattern mining algorithms. For visual representation of big data, the behaviour of Internet users was analyzed using geometric methods, multidimensional scaling and artificial neural networks.

The research completed in this thesis showed the following conclusions:

1) Approximate (probabilistic) frequent sequence mining methods could deliver the acceptable results with right balance between the computation speed and precision when dealing with significantly growing amount of digital information in big data; whereas the exact frequent sequence mining methods work under the time constraints and cannot guarantee the computation speed required by various real-world applications.
2) Novel approximate RSM method brings a significant benefit among existing approximate frequent sequence mining methods by providing a theoretical estimation of error probabilities. It is proved that the error probabilities made by this method could be theoretically estimated using standard statistical methods. The theoretical estimation allows using the proposed method without any extensive empirical experiments, whereas other state-of-the-art approximate methods provide only the empirical evidence based on extensive experiments and observations of algorithm results on different databases.
3) The performance studies demonstrated that a novel approximate MRM method proposed in this thesis has shown the best overall result in terms of the precision among other approximate algorithms, but it was outperformed by RSM method in terms of the computation speed. If the recommendations on the RSM algorithm parameters are followed, then even more precise results could be achieved.

4) The approximate frequent pattern mining algorithms that require reading the entire original database do not satisfy the time constraints required by many real-world applications. Therefore MPBM method proposed in this thesis does not satisfy the time constraints as it requires reading the original database at least several times (the number of times equals to the order of the Markov process) and it delivers similar precision results as MRM method.
5) In visual representation of big data, SOM visualization showed the best results when visualizing the Internet users' behaviour; however it requires some training and experience by a user, who is analyzing the data. Based on the visualization, the decisions on the advertising strategy could be taken by a non-technical person who is managing the advertising campaigns.