# Turinys



# ĮVADAS

Duomenų[DHS] kiekis pasaulyje labai sparčiai didėja. Dar daugiau, tie duomenys sudėtingėja. Duomenų sudėtingėjimas suponuoja atributų skaičiaus augimą - didėja duomenų dimensijų skaičius. Dimensijų skaičiaus „sprogimas“ ypač pastebimas biologiniuose duomenyse. Taip yra todėl, nes atsiranda vis naujesni būdai apdoroti biologinius duomenis, pavyzdžiui, genomo sekvenavimas ar epigenitinė analizė. Sekvenuojant genomą moderniomis priemonėmis galima gauti net keletą milijonų dimensijų vienam genui(!). Todėl naujiems duomenims klasifikuoti reikia ir naujų klasifikavimo strategijų, nes tradicinės yra tiesiog nepajėgios dirbti su tokiu milžinišku dimensijų skaičiumi - drastiškai didėja skaičiavimo laikas, mažėja klasifikavimo tikslumas.

Viena iš strategijų dirbti su didelio dimensiškumo duomenis yra naudoti įvairius dimensijų atrinkimo[GE03] (angl. feature selection) metodus. Tokie metodai nėra naujiena, jie jau kuris laikas yra naudojami mašininio mokymosi taikymuose, bet išradingai naudojant jie gali itin svariai prisidėti prie naujų klasifikavimo strategijų, reikalingų darbui su daugiamačiais duomenimis. Dimensijų atrinkimo etapas yra svarbus duomenų paruošimui. Jis ne tik leidžia sukurti tikslesnius klasifikavimo modelius, taip pat jis padeda geriau suprasti ir vizualizuoti tais duomenimis apibūdinamus procesus, sumažinti atminties

poreikį, pagreitinti mokymosi ir sumažinti klasifikatorių darbo laiką. Tačiau ir dimensijų atrinkimo metodai turi būti pakankamai našūs, todėl ne visi yra naudotini.

Gyvybėms mokslų tyrėjams itin svarbu fokusuoti į mažesnį dimensijų skaičių, nes tai itin paspartina jų tyrimus - jiems reikia tirti mažesnį skaičių mėginių. Mažesnio skaičiaus mėginių tyrimas ir kainuoja mažiau, nes mažiau reikia ir cheminių reagentų ir darbo laiko. Bet čia iškyla naujas dimensijų atrinkimo kriterijus - stabilumas (ang. robustness). Stabilumas didina tikimybę, kad atrinktosios dimensijos yra tikrai susijusios su nagrinėjama problema. Jei dimensijų atrinkimo rezultatai labai stipriai varijuos, tai tada reikės atlikinėti bandymus su visa duomenų aibe, kas yra labai neefektyvu. Pastebėtina, kad stabilumo matavimai turi būti atliekami būtinai atsižvelgiant į klasifikavimo tikslumą.

Dar viena problema dirbant su daugiamačiais duomenimis yra tai, kad dažnai turimas labai ribotas skaičius duomenų objektų. Dimensijų - objektų santykis neretai skiriasi visomis eilėmis. Atrodytų, tik laiko klausimas, kada bus paruošta daugiau objektų, bet žvelgiant į duomenų gavybos tendencijas pasidaro aišku, kad objektų skaičius niekada nepavys dimensijų skaičiaus. Todėl reikia labai apgalvoti, kaip bus į tai atsižvelgta kuriant klasifikavimo modelius. Nes esant mažam objektų skaičiui kyla grėsmė susidurti su persimokymo problema (angl. overfitting). Tokiu atveju klasifikatorius tampa bevertis.

Duomenyse daugėja dimensijų, todėl neišvengiamai daugėja ir triukšmo. Triukšmas atsiranda dėl įvairių priežasčių, pavyzdžiui, cheminiai preparatai buvo ne visai tinkamai paruošti. Triukšmas duoda atsitiktinius rezultatus, iš kurių naudos nedaug. Atsiranda poreikis identifikuoti triukšmingus duomenis ir juos išmesti iš klasifikavimo proceso.

Taigi, žvelgiant iš paukščio skrydžio, galime pastebėti, kad norint optimaliai dirbti su daugiamačiais duomenimis vienu metu reikia atsižvelgti į eilę kriterijų:

1. Klasifikavimo tikslumas - tipinė klasifikavimo užduotis yra atskirti sergančius pacientus nuo sveikų.;
2. Dimensijų atrinkimo stabilumas, atsižvelgiant į klasifikavimo rezultatus;
3. Didelis dimensijų-objektų santykis;
4. Triukšmo lygis duomenyse;
5. Skaičiavimo išteklių optimalus naudojimas.

Reikalavimas vienu metu atsižvelgti į keletą kriterijų labai apsunkina užduotį. Daugiamačių duomenų klasifikavime pagrindinė problema yra surasti optimalų metodą, kuris geriausiai atsižvelgia į aukščiau minėtus kriterijus.

Šiame darbe dirbsime su viešai prieinamomais genų ekspresijos duomenų rinkiniais.

1 lentelė. Darbe naudoti duomenų rinkiniai

| **Pavadinimas** | **Šaltinis** | **Objektų skaičius (+/-)** | **Dimensijų skaičius** | **ODS** |
|---|---|---|---|---|
| Gaubtinės žarnos auglys (angl. Colon) | [ABN+99] | 62 (40/22) | 2000 | 3,1% |
| Centrinės nervų sistemos auglys (CNS) | [PTG+02] | 60 (39 AAL / 21 AML) | 7129 | 0.84% |
| Prostatos auglys | [SFR+02] | 102 (52/50) | 6033 | 1.7% |
| Šizofrenija ir maniakinė depresija | [Ins] | 90 (bp[1]: sz[2]: cc[3] =30:31:29) | 22283 | 0.403% |

Duomenų rinkinius apibūdinantis dydis ODS[4], kuris turimiems duomenims tesiekia nuo 0,403% iki 3,01% procento, parodo, kad turime labai retus (angl. scarce) duomenis, o tai labai apsunkina mokymosi procesą ir gali sukelti persimokymo (angl. overfitting) problemą.

Skaičiavimo išteklius, kurių reikėjo ne taip ir mažai, suteikė VU MIF ITTC. Programavimo darbai buvo atlikti naudojant R programavimo kalbą. Didžiąją dalį eksperimentų atlikau profesinės praktikos MII metu.

Su teorine dalykinės srities medžiaga susipažinau skaitydamas darbo vadovo rekomenduotus mokslinius straipsnius, bei naudodamasis internetine paieška.

Šio darbo tikslas yra išanalizuoti darbo su daugiamačiais duomenis ypatybes.

Šiam darbui yra keliamos tokios užduotys:

1. Apžvelgti esamus klasifikavimo metodus;
2. Išanalizuoti populiariausių dimensijų atrinkimo metodų spartą;
3. Išanalizuoti populiariausių dimensijų atrinkimo metodų stabilumą;
4. Išanalizuoti kaip dimensijų atrinkimo metodai įtakoja klasifikatorių tikslumą.

# 1. DIMENSIJŲ ATRINKIMAS

Klasifikuojant daugiamačius duomenis susiduriame su taip vadinamu dimensiškumo prakeiksmu (angl. curse of dimentionality). Pavyzdžiui, kai di-

[1]bp (angl. Bipolar disorder) - maniakine depresija sergantys pacientai.
[2]sz (angl. Schizophrenia) - šizofrenija sergantys pacientai.
[3]cc (angl. Control Crowd) - kontrolinė grupė.
[4]ODS - Objektų dimensijų santykis

mensijų skaičius įkopia į trečią ar ketvirtą eilę naudoti paprastus klasifikavimo algoritmus tampa nebeefektyvu nei laiko, nei klasifikavimo našumo atžvilgiais. Vienas iš būdų kovoti su dimensiškumo prakeiksmu yra naudoti vienokius ar kitokius dimensijų skaičiaus mažinimo metodus. Šiame skyriuje nagrinėsiu bazinius dimensijų atsirinkimo metodus, keletos kriterijų suliejimą metodą (angl. feature selection based on multicriterion fusion)[YM11], bei stabilių dimensijų grupių išskyrimo metodą[LYD09].

## 1.1. Baziniai dimensijų atrinkimo metodai

Yra pasiūlyta daugybė dimensijų atrinkimo metodų. Šiame skyriuje aptarsiu keletą taip vadinamų bazinių[5] dimensijų atrinkimo metodų:

1. Fisher'io įvertis (angl. Fisher ratio)[PWCG01];
2. Atpalaidavimo (ang. relief) metodas[RSK03];
3. Asimetrinis priklausomybės koeficientas[Sha01] (ADC) (angl. Asymmetric Dependency Coefficient);
4. Absoliučių svorių SVM[Vap00] (AW-SVM) (angl. Absolute Weight SVM)
5. Rekursyvus dimensijų eliminavimas pagal SVM[GWBV02] (SVM-RFE) (angl. Recursive Feature Elimination by SVM)

### 1.1.1. Fisher'io įvertis

Fisher'io įvertis vertina individualias dimensijas pagal jų klasių atskiriamąją galią. Dimensijos įvertis yra sudarytas iš tarpklasinio skirtumo santykio su vidiniu klasės pasiskirstymu:

$$FR(j) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \tag{1}$$

kur, $j$ - yra dimensijos indeksas, $\mu_{jc}$ - dimensijos $j$ reikšmių vidurkis klasėje $c$, $\sigma_{jc}^2$ - dimensijos $j$ reikšmių standartinis nuokrypis klasėje $c$, kur $c = 1, 2$. Kuo didesnis yra Fisher'io įvertis, tuo geriau ta dimensija atskiria klases.

### 1.1.2. Atpalaidavimo metodas

Atpalaidavimo metodas iteratyviai skaičiuoja dimensijų „susietumą“. Pradžioje „susietumas“ visoms dimensijoms yra lygus nuliui. Kiekvienoje iteracijoje

[5] Aptarsiu tik bazinius todėl, kad jie yra svarbiausi, nes jų tarpusavio rezultatai yra nekoreliuojantys, o nekoreliuojančius rezultatus galima panaudot juos apjungiant, taip gauname sinergijos efektą.

atsitiktinai[6] pasirenkamas objektas iš duomenų bazės, surandami artimiausi kaimynai iš tos pačios ir kitos klasės, ir atnaujinamos visų dimensijų „susietumo“ reikšmės. Dimensijos įvertis yra vidurkis visų objektų atstumų iki artimiausių kaimynų iš tos pačios ir kitos kasės:

$$W(j) = W(j) - \frac{diff(j, x, x_H)}{n} + \frac{diff(i, x, x_M)}{n}, \tag{2}$$

kur $W(j)$ - $j$-osios dimensijos „susietumo“ įvertis, $n$ - objektų aibės dydis,$x$ - atsitiktinai pasirinktas objektas, $x_H$ - artimiausias kaimynas iš tos pačios klasės (angl. nearest-Hit), $x_M$ - artimiausias kaimynas iš kitos klasės, $diff(j, x, x')$ - $j$-osios dimensijos reikšmių skirtumas tarp laisvai pasirinkto objekto $x$ ir atitinkamo kaimyno, kur skirtumą į intervalą $[0, 1]$ normalizuojanti funkcija yra:

$$diff(j, x, x') = \frac{|x_j - x'_j|}{x_{jmax} - x_{imin}}, \tag{3}$$

kur $x_{jmax}$ ir $x_{jmin}$ yra maximali ir minimali $j$-osios dimensijos reikšmės. „Susietumo“ reikšmių atnaujinimas yra vykdomas $n$ kartų ir kuo didesnė galutinė reikšmė, tuo svarbesnė dimensija. Pastebėtina, kad šis algoritmas veikia tik su dviejomis klasėmis, nors yra ir išplėtimų.

### 1.1.3. Asimetrinis priklausomybės koeficientas

Asimetrinis priklausomybės koeficientas yra dimensijų reitingavimo motodas, kuris matuoja klasės $Y$ etiketės (angl. label) tikimybinę priklausomybę $j$-ąjai dimensijąi, naudodamas informacijos prieaugį (angl. information gain):

$$ADC(Y, j) = \frac{MI(Y, X_j)}{H(Y)}, \tag{4}$$

kur $H(Y)$ - klasės $Y$ entropija, o $MI(Y, X_j)$ - yra bendrumo informacija (angl. mutual information) tarp klasės etiketės $Y$ ir $j$-osios dimensijos

$$H(Y) = -\sum_y p(Y = y) log p(Y = y), \tag{5}$$

$$H(X_j) = -\sum_x p(X_j = x) log p(X_j = x), \tag{6}$$

$$MI(Y, X_j) = H(Y) + H(X_j) - H(Y, X_j), \tag{7}$$

$$H(Y, X_j) = -\sum_{y, x_j} p(y, x_j) log p(y, x_j), \tag{8}$$

Kuo didesni ADC įverčiai, tuo dimensija yra svarbesnė, nes turi daugiau informacijos apie duomenų klases.

[6] Pastebėtina, kad metodas turi atsitiktinį elementą, todėl klasifikavimo ir dimensijų atrinkimo stabilumo rezultatai dažniausiai šiek tiek varijuoja nekeičiant konfigūracijos.

### 1.1.4. Absoliučių svorių SVM

Atraminių vektorių metodas (SVM) yra vienas populiariausių klasifikavimo algortimų, nes jis gerai susidoroja su daugiamačiais duomenimis. Yra keletas bazinių SVM variantų, bet šiame darbe naudosime tiesinį SVM, nes jis demonstruoja gerus rezultatus dirbant su genų ekspresijos duomenimis. Tiesinis SVM yra hiperplokštuma apibrėžta kaip:

$$\sum_{j=1}^{p} w_j x_j + b_0 = 0, \tag{9}$$

kur $p$ - dimensijų kiekis, $w_j$ - j-osios dimensijos svoris, $x_j$ - j-osios dimensijos kintamasis, $b_0$ - konstanta. Dimensijos absoliutus[7] svoris $w_j$ gali būti panaudotas dimensijų reitingavimui. Pastebėtina, kad svorių nustatymas yra atliekamas tik vieną kartą[8].

### 1.1.5. Rekursyvus dimensijų eliminavimas pagal SVM

Rekursyvus dimensijų eliminavimas pagal SVM yra vienas populiariausių dimensijų atrinkimo algoritmų. Todėl, jis yra naudojamas, kaip atskaitos taškas (angl. threshold) vertinant kitus dimensijų atrankos metodus. Iš esmės šis metodas yra daugkartinis absoliučių svorių SVM metodo taikymas nuolat išmetinėjant dimensijas su mažiausiais svoriais. Rekursyvus dimensijų eliminavimas mums padeda surasti tinkamą dimensijų poaibį, kas nevisada pavyksta su dimensijų reitingavimo metodais. Bendroji rekursyvaus dimensijų eliminavimo procedūra: Jei trečiajame algoritmo žingsnyje yra pašalinama tik viena

**Algoritmas nr. 1** Rekursyvus dimensijų eliminavimas

1. Turime pilną dimensijų rinkinį $F_0$, nustatome $i = 0$.
2. Įvertiname kiekvienos dimensijos kokybę dimensijų aibėje $F_i$.
3. Išmetame mažiausiai svarbią dimensiją iš $F_I$ tam, kad gautume dimensijų rinkinį $F_{i+1}$.
4. Nustatome $i = i + 1$ ir grįžtame į antrąjį žingsnį kol nėra patenkinta algoritmo pabaigos sąlyga.

dimensija, tai gauname dar ir dimensijų reitingavimą, o jei pašalinamos kelios dimensijos ar jų dalis (pvz. 50%) tai reitingavimo negauname. Pastebėtina, kad rekursyvus dimensijų eliminavimas gali labai padidinti algoritmo sudėtingumą skaičiavimo resursų atžvilgiu. Algoritmo pabaigos sąlyga gali būti koks nors konkretus dimensijų skaičius arba tiesiog dimensijų aibę mažinti tol, kol dimensijų visai nebeliks.

[7] Svorį reikia imti absoliutaus dydžio, nes neigiamas svoris implikuoja priklausomybę vienai klasei, o teigiamas kitai klasei.

[8] SVM-RFE - metodas svorius nustato daug kartų.

## 1.2. Multikriterinis dimensijų atrinkimas

Multikriterinio dimensijų atrinkimo metodų esmė yra panaudoti kelis dimensijų atrinkimo metodus suliejant jų rezultatus į vieną bendrą rezultatą. Kodėl naudinga sulieti keletą dimensijų atrinkimo metodų rezultatų? Sulieti keletą dimensijų atrinkimo metodų rezultatų naudinga, nes pavieniai dimensijų atrinkimo metodai be to, kad turi savitų privalumų, visada turi ir savo silnybių, pavyzdžiui, jautrumas išimtims (angl. outliers), negali rasti dimensijų tarpusavio priklausomybių, etc. Yra skiriamos trys priežastys, kodėl keletas kombinuotų silpnų ir nestabilių dimensijų atrinkimo metodų gali duoti geresnius rezultatus[Die00]:

1. Keletas skirtingų, bet vienodai optimalių hipotezių gali būti teisingos, ir kriterijų kombinavimas sumažiną tikimybę, kad bus pasirinkta neteisinga hipotezė;
2. Atskiri kriterijai gali dirbti skirtinguose lokaliuose optimumuose, tuo tarpu kombinavimas gali geriau reprezentuoti tikrąją duomenų funkciją;
3. Tikroji duomenų funkcija negali būti reprezentuojama jokia hipoteze paskiro algoritmo hipotezių erdvėje ir agreguojant pavienių metodų rezultatus galima praplėsti hipotezių erdvę.

Suliejant keletą skirtingų metodų suliejamos gerosios pavienių dimensijų atrinkimo metodų savybės, taip kompensuojant algoritmų silpnybes.

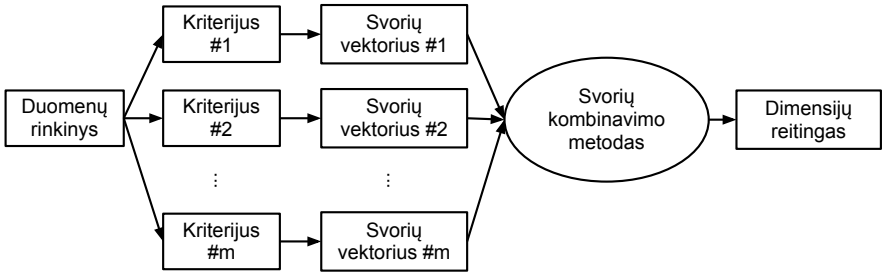
Galima pavienių dimensijų atrankos rezultatus sulieti pagal šias suliejimo strategijas:

1. Svoriais grįstas multikriterinis suliejimas;
2. Reitingais (angl. rank) grįstas multikriterinis suliejimas;
3. Svoriais ir reitingais grįstas multikriterinis suliejimas.

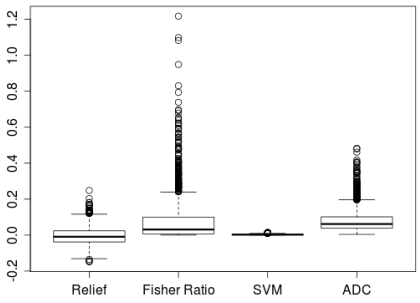
### 1.2.1. Svoriais grįstas multikriterinis suliejimas

Svoriais grįsto multikriterinio dimensijų atrinkimo suliejimo pagal svorius algoritmo pirmajame žingsnyje kiekvienas bazinis metodas priskiria duomenų rinkinio dimensijoms svorius, tada tie svoriai yra kombinuojami į vieną sutarties (angl. consensus) svorių vektorių, kurio pagrindu yra gaunami dimensijų reitingai. Algoritmas yra pavaizduotas 1 pav.

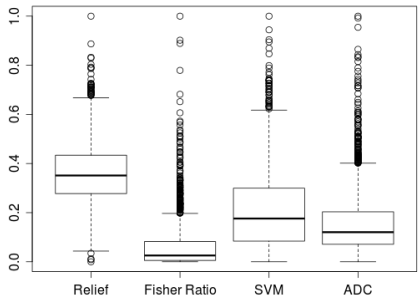
Suliejant svorius svarbu yra užtikrinti, kad svoriai, gauti naudojant skirtingus bazinius kriterijus, būtų palyginami. Todėl svorių normalizavimas turi būti atliekamas prieš svorių kombinavimą. Kitu atveju dimensijų įvertinimo metodai bus nepalyginami. Paveikslėlyje 2 pav. nenormalizuotų pavienių dimensijų vertinimo metodų skiriasi netgi suteiktų svorių intervalai. Paveikslėlyje 3 pav. matome, kad net ir normalizavus svorius gana stipriai skiriasi svorių




1 pav.: Svoriais grįstas multikriterinis suliejimas.




2 pav.: Pavienių dimensijų atrinkimo metodų nenormalizuotas svorių pasiskirstymas.




3 pav.: Pavienių dimensijų atrinkimo metodų normalizuotas svorių pasiskirstymas.

kvartiliai - į tai reikia atkreipti dėmesį interpretuojant galutinius dimensijų vertinimo rezultatus. Šiame darbe svoriai yra normalizuoti intervale $[0, 1]$ pagal formulę:

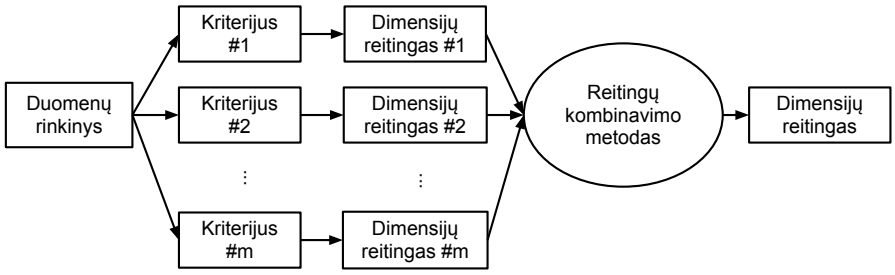
$$u'_i = \frac{u_i - u_{imin}}{u_{imax} - u_{imin}}, \tag{10}$$

kur $u_i$ - dimensijų svorių vektorius pagal $i$ kriterijų, $u_{imin}$ - minimali $u_i$ svorių vektoriaus reikšmė, $u_{imax}$ - maksimali $u_i$ svorių vektoriaus reikšmė, $u'_i$ - normalizuotų svorių vektorius.

Sutarties svorių vektorius $u$ yra vidurkis normalizuotų svorių vektorių:

$$u = \frac{1}{m} \sum_{i=1}^{m} u'_i, \tag{11}$$

kur $m$ yra bazinių kriterijų skaičius. Reikia paminėti, kad didesnė svorio reikšmė reiškia, kad dimensija yra geresnė.




4 pav.: Reitingais grįstas multikriterinis suliejimas.

### 1.2.2. Reitingais grįstas multikriterinis suliejimas

Reitingais grįsto multikriterinio suliejimo pagal reitingus metodas gauna duomenų rinkinio dimensijų reitingą, pagal keletą bazinių dimensijų reitingavimo kriterijų. Algoritmo pirmajame žingsnyje keletas dimensijų atrinkimo kriterijų grąžina dimensijų reitingu, paskui tie reitingai yra kombinuojami į vieną bendra dimensijų reitingą. Algoritmas yra pavaizduotas 4 pav. Suliejimo pagal reitingus metodas nereikalauja dimensijų atrinkimo metodų rezultatų normalizavimo, nes tiesiog imame dimensijoms priskirtus reitingus ir juos kombinuojame. Skirtingai nei suliejimo pagal svorius algoritme, baziniai dimensijų atrinkimo kriterija dimensijų eliminavimas[YM11] susideda iš dviejų dalių: keletos dimensijų atrinkimi turi gražinti dimensijų reitingus, o ne svorius.

Dimensijų reitingų kombinavimui yra keletas metodų[DKNS01], tačiau paprastumo dėlei šiame darbe naudosiu Borda balsavimą[9] (angl. Borda count). Tarkime, kad turime $m$ basuotojų ir $p$ kandidatų aibę. Tada Borda balsavimo metodas kiekvienam $i$-ajam balsuotojui sukuria balsų vektorių $v_i$ tokiu būdu: geriausiai įvertintam kandidatui suteikiama $p$ taškų, antrajam kandidatui $p-1$, ir t.t. Galutiniai taškai yra gaunami sudedant visų balsuotojų taškus

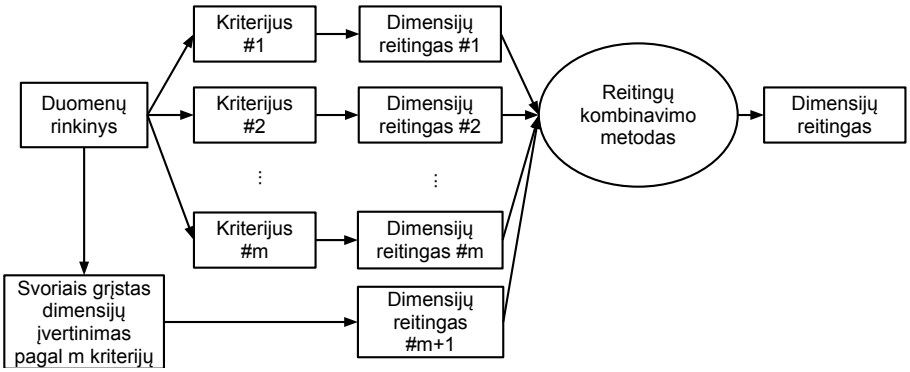
$$v = \sum_{i=1}^{m} v_i, \tag{12}$$

kur $v$ yra suminių taškų vektorius, o iš jo galime gauti ir dimensijų reitingus.

### 1.2.3. Svoriais ir reitingais grįstas multikriterinis suliejimas

Svoriais ir reitingais grįsto multikriterinio suliejimo metodas nuo reitingais grįsto multikriterinio suliejimo metodo skiriasi tuo, kad kaip dar vienas reitingas yra panaudojamas svoriais grįsto multikriterinio dimensijų atrinkimo metu gautas reitingas. Multikriterinio dimensijų įverčių ir pagal svorius, ir pagal reitingus metodas vyksta trimis žingsniais:

[9] Dar žinomas kaip „Pažymių metodas“. Jis buvo pasiūlytas prancūzų matematiko ir fiziko Jean-Charles de Borda 1770 metais.




5 pav.: Svoriais ir reitingais grįstas multikriterinis suliejimas.

1. Gauname dimensijų reitingus pagal $m$ pavienių dimensijų atrinkimo motodų;
2. Suliejame dimensijų įverčius pagal svorius ir taip gauname vieną dimensijų reitingą;
3. Reitinguojame dimensijas pagal visus turimus $m + 1$ pavienius reitingus.

Algoritmas yra pavaizduotas 5 pav.

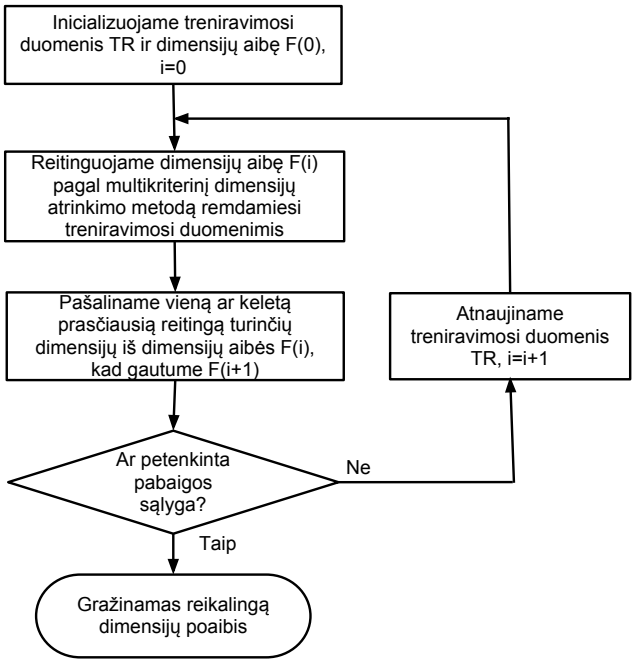
Kadangi yra suliejami keli mažai koreliuojantys dimensijų reitingavimo metodai, yra pasiekiamas didesnis dimensijų atrinkimo stabilumas, kai varijuoja treniravimosi duomenų poaibis (angl. subsampling).

### 1.2.4. Multikriterinis rekursyvus dimensijų eliminavimas

Jei dimensijų atrinkimo tikslas yra pagerinti klasifikavimo rezultatus, tai taikymas multikriterinio dimensijų atrinkimo metodų nebūtinai duos pageidaujamą rezultatą, nes yra pastebėta, kad vien dimensijų reitingavimas nebūtinai suranda geriausią dimensijų poaibį. Tam, kad būtų surastas geriausias dimensijų poaibis reikia kombinuoti multikriterinį dimensijų reitingavimą su paieškos strategija. Rekursyvus dimensijų eliminavimas yra dažnai naudojama paieškos strategija dimensijų atrinkimui. Todėl yra kombinuojamas multikriterinis dimensijų reitingavimas ir rekursyvus dimensijų eliminavimas.

Multikriterinis rekursyvus dimensijų eliminavimas[YM11] susideda iš dviejų dalių: keletos dimensijų atrinkimo kriterijų suliejimo ir pagal svorius, ir pagal reitingus, ir rekursyvaus dimensijų eliminavimo aprašyto algoritme nr. 1. Algoritmas pavaizduotas 6 pav.

Yra pastebėta, kad standartinis rekursyvus dimensijų eliminavimas, kai vienos iteracijos metu yra eliminuojama viena dimensija, gali labai padidinti




6 pav.: Multikriterinio rekursyvaus dimensijų eliminavimo algoritmas.

algoritmo sudėtingumą. Todėl genų ekspresijos duomenims yra rekomenduotina eliminuoti keletą dimensijų vienu metu.

Nors SVM-RFE dimensijų atrinkimo algoritmas ir yra labai populiarus, tačiau yra žinoma, kad jam trūksta stabilumo. Todėl kombinuodami didesnį stabilumą turintį multikriterinį dimensijų atrinkimą su rekursyvaus dimensijų eliminavimo paieškos strategija, turėtume gauti stabilesnį dimensijų atrinkimo algoritmą.

# 2. DIMENSIJŲ ATRINKIMO STABILUMAS

Dimensijų atrinkimo technikų stabilumas yra vis didesnę svarbą įgaunanti tyrimų kryptis. Stabilumo aktualumas yra sąlygotas to, kad biologiniuose duomenyse galima gana užtikrintai daryti prielaidą, kad konkrečiai problemai yra aktualios tik tam tikros dimensijos. Todėl dalykinės srities ekspertams yra aktualu naudoti tik tuos dimensijų atrinkimo metodus, kurie yra stabilūs ir relevantiški modeliuojamai problemai.

Šiame skyriuje apžvelgsime teorinį stabilumo matavimų modelį. Taip pat įvertinsime pavienių dimensijų atrinkimo metodų stabilumą taikant juos

įvairioms duomenų bazėms. Taip pat išanalizuosime situaciją, kai kombinuojami keletos dimensijų atrinkimo metodų rezultatai.

## 2.1. Stabilumo matavimas

Vertinant dimensijų atrinkimo metodų stabilumą yra svarbu kaip panašiai yra atrenkamos dimensijos, kai yra atliekamas dimensijų atrinkimas su vis kitu duomenų poaibiu. Kuo panašesnius dimensijų atrinkimo rezultatus gauname, tuo stabilumas yra didesnis. Vidutinis stabilumas gali būti apibrėžtas kaip vidurkis visų reitingavimo metu gautų sąrašų porų tarpusavio panašumo įverčių:

$$S_{tot} = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} S(f_i, f_j)}{k * (k-1)}, \tag{13}$$

kur $k$ žymi kiek kartų buvo imtas skirtingas poaibis objektų dimensijų atrinkimui, $f_i$, $f_j$ - dimensijų atrinkimo rezultatas - reitingai, $S(f_i, f_j)$ - yra kokia nors panašumo matavimo funkcija.

Kaip matome dimensijų atrinkimo stabilumas priklauso nuo to, kokią panašumo funkciją naudosime. Tradicinės panašumo funkcijos (persidengimo procentas, Pearson'o koreliacija, Spearman'o koreliacijoa, Jaccard indeksas) gali būti taikomos, bet jos yra linkusios priskirti didesnes panašumo reikšmes, kai pasirenkamas didesnis dimensijų poaibis. Taip yra dėl padidėjusio sisteminio nuokrypio (ang. bias), nes imant didesnį poaibį padidėja tikimybė tiesiog atsitiktinai pasirinkti dimensiją. Kad išventume šios problemos panašumui vertinti buvo pasirinktas Kunchevos [Kun07] indeksas:

$$KI(f_i, f_j) = \frac{r * N - s^2}{s * (N - s)} = \frac{r - (s^2/N)}{s - (s^2/N)}, \tag{14}$$

kur $s = |f_i| = |f_j|$ yra atrinktų dimensijų aibės dydis, $r = |f_i \bigcap f_j|$ - abiems atrinktiems dimensijų poaibiams bendrų dimensijų skaičius, $N$ - bendras duomenų aibės dimensijų skaičius. Pastebėtina, kad formulėje esantis atėminys $s^2/N$ ištaiso sisteminį nuokrypį atsirandantį dėl galimybės atsitiktinai pasirinkti dimensijas. Kunchevos indeksas gali įgyti reikšmes iš intervalo $[-1, 1]$, kur didesnė reikšmė reiškia didesnį panašumą, o artimos nuliui reikšmės reiškia, kad dimensijos atrenkamos daugiausia atsitiktinai. Kunchevos indekso ypatybė yra ta, kad jis atsižvelgia tik į persidengiančias, tačiau visiškai nekreipia dėmesio į koreliuojančias dimensijas.

Vertinant stabilumą tarp skirtingų metodų gali iškilti problemų, nes ne visi dimensijų atrinkimo metodai gražina rezultatą tokiu pačiu formatu. Šiame darbe dimensijų atrinkimo metodų rezultatas yra ne dimensijai priskirtas svoris, bet dimensijos reitingas. Todėl $f_i$ yra sąrašas, kurio ilgis yra $N$, kur pirmas sąrašo elementas yra geriausią reitingą turinčios dimensijos numeris, o paskutinis sąrašo elementas yra blogiausią reitingą turinčios dimensijos numeris.

Galiausiai yra svarbu paminėti, kad dimensijų stabilumas nėra matuojamas visiškai nepriklausomai - jis yra matuojamas atsižvelgiant į klasifikavimo rezultatus. Dimensijų atrinkimo metodų stabilumas yra matuojamas tik tada kai atrinktos dimensijos duoda gerus klasifikavimo rezultatus. Taip yra, nes kokios nors dalykinės srities ekspertui, nėra naudingos tos dimensijų atrinkimo strategijos, kurios duoda labai stabilius rezultatus, bet nėra naudingos klasifikavimo modelio kūrime.

## 2.2. Pavienių dimensijų atrinkimo metodų stabilumas

Šiame skyriuje apžvelgsime pavienių dimensijų atrinkimo motodų stabilumą. Stabilumas visiems metodams buvo matuojamas atsižvelgiant į klasifikavimo rezultatus - buvo matuojamas stabilumas tikra

### 2.2.1. Fisher'io dimensijų atrinkimo metodas

# Literatūra


[ABN+99] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999.

[DHS] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. `http://books.google.lt/books?id=YoxQAAAAMAAJ`.

[Die00] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited. In *Proceedings of WWW10*, pages 613–622, 2001.

[GE03] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[GWBV02] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 2002.

[Ins] Stanley Medical Research Institute. Online genomics database. [žiūrėta 2012-04-03].

[Kun07] Ludmila I. Kuncheva. A stability index for feature selection. In *Artificial Intelligence and Applications*, pages 421–427, 2007.

[LYD09] Steven Loscalzo, Lei Yu, and Chris Ding. Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 567–576, New York, NY, USA, 2009. ACM.

[PTG+02] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[PWCG01] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on Computational biology*, RECOMB '01, pages 249–255, New York, NY, USA, 2001. ACM.

[RSK03] Marko Robnik-Sikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69, 2003.

[SFR+02] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[Sha01] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[Vap00] V.N. Vapnik. *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, 2000.

[YM11] F. Yang and KZ Mao. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(4):1080–1092, 2011.