KAUNO TECHNOLOGIJOS UNIVERSITETAS

FUNDAMENTALIŲJŲ MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

Simonas Šimkevičius

KLASIFIKAVIMO SU MOKYTOJU METODŲ LYGINAMOJI ANALIZĖ

Magistro darbas

Darbo vadovas

doc. dr. V. Janilionis

Kaunas, 2006

KAUNO TECHNOLOGIJOS UNIVERSITETAS

FUNDAMENTALIŲJŲ MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

TVIRTINU
Katedros vedėjas
prof. dr. J. Rimas
2006-05

KLASIFIKAVIMO SU MOKYTOJU METODŲ LYGINAMOJI ANALIZĖ

Taikomosios matematikos magistro baigiamasis darbas

Kalbos konsultantė
Lietuvių k. katedros lekt. dr.
J. Džežulskienė
2006-05

Vadovas
doc. dr. V. Janilionis
2006-05

Recenzentas
2006-05

Atliko
FMMM-4 gr. stud.
S. Šimkevičius
2006-05-16

Kaunas, 2006

KVALIFIKCINĖ KOMISIJA

Pirmininkas: Leonas Saulis, profesorius (VGTU)

Sekretorius: Eimutis Valakevičius, docentas (KTU)

Nariai: Algimantas Jonas Aksomaitis, profesorius (KTU)

Vytautas Janilionis, docentas (KTU)

Vidmantas Povilas Pekarskas, profesorius (KTU)

Rimantas Rudzkis, profesorius (MII)

Zenonas Navickas, profesorius (KTU)

Arūnas Barauskas, UAB „Elsis“ generalinio direktoriaus pavaduotojas

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

Šimkevičius S. A comparative analysis of supervised classification methods : Master‘s work in applied mathematics / supervisor dr. assoc. prof. V. Janilionis; Department of Applied mathematics, Faculty of Fundamental Sciences, Kaunas University of Technology. – Kaunas, 2006. – 101 p.

SUMMARY

Supervised classification methods are applied in many fields. The main problem of applying these methods is how to select the most appropriate method in particular case. The literary review was fulfilled and the advantages and disadvantages of mostly used criterion of supervised classification methods comparisons were ascertained. Then the methodology of comparisons was suggested. The analysis of SAS system procedures and macro commands was made. It was ascertained that there is not comfortable software which allows comparing the results of supervised classification methods. This work demands a lot of work, good knowledge of SAS programming language and high qualification in programming. So, the main purpose of this work is to expand the statistical data analysis system SAS possibilities in comparison of supervised classification methods and classificate various data.

In this work the possibilities of SAS system are expanded by the tool which allows comparing quality of the linear, quadratic, kernel, nearest neighbor’s discriminant analysis and logistic regression analysis methods. There were used classification error estimates which were got by resubstition, cross–validation leave one out, bootstrap and Monte Carl cross–validation methods, although classification error confidence intervals which were got by non-parametric bootstrap method. The test of created tool was made with various data (different sample sizes, various classis separability, violations of assumptions about the underlined data and etc.). The most appropriate method of classification was selected and quality of classification was estimated.

ĮVADAS

Klasifikavimo metodai taikomi daugelyje sričių, pvz., medicinoje – ligoms diagnozuoti, draudimo srityje – klientams skirstyti į įvairias rizikos grupes. Visi klasifikavimo metodai skirstomi į dvi pagrindines klases – klasifikavimą su mokytoju ir be mokytojo. Darbe analizuoti tik pirmieji (tiesinė, kvadratinė, branduolinė, artimiausių kaimynų diskriminantinė analizė, logistinė regresinė analizė, neuroniniai tinklai, klasifikavimo medžiai), kurie taikomi tada, kai egzistuoja mokomoji imtis, t.y. informacija apie dalies tiriamų objektų požymių reikšmes ir priklausomybę konkrečiai klasei.

Klasifikavimo su mokytoju metodų sukurta daug. Tiems patiems duomenims taikant skirtingus metodus, dažnai gaunami skirtingi rezultatai. Daug euristinių, teorinio pagrindimo neturinčių, metodų. Todėl pagrindinė šių metodų taikymo problema – kaip konkrečiu atveju parinkti tinkamiausią metodą.

Atlikta klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo literatūros apžvalga parodė, kad nėra universalių metodų ir patogių programinių priemonių leidžiančių kiekvienu atveju parinkti tinkamiausią metodą. Net klasifikavimo su mokytoju metodų taikymo rezultatų lyginimui naudojant vieną galingiausių statistinės duomenų analizė sistemų SAS, reikia didelių darbo sąnaudų ir gerų SAS programavimo kalbos žinių bei įgūdžių. Todėl pagrindinis darbo tikslas - išplėsti SAS sistemos galimybes klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankiu ir atlikti pasirinktų duomenų klasifikavimą.

Darbe apžvelgti naudojami klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo kriterijai, nustatyti jų trūkumai ir privalumai bei pasiūlyta klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo metodika. Klasifikavimo metodų taikymo rezultatai lyginami naudojant keturis klaidingo klasifikavimo tikimybės taškinius įverčius, gautus savos imties, kryžminio patikrinimo, įkelčių bei Monte Karlo kryžminio patikrinimo metodais. Siekiant atsakyti į klausimą, ar skirtingų metodų klasifikavimo kokybė statistiškai reikšmingai skiriasi, naudojami klaidingo klasifikavimo tikimybės intervaliniai įverčiai, gauti neparametriniu įkelčių metodu. Darbe tikrinamos parametrinių klasifikavimo su mokytoju metodų taikymo prielaidos, t.y. tikrinama ar diskriminavimo kintamųjų skirstinys yra daugiamatis normalusis ir ar skirtingose klasėse kovariacijų matricos identiškos.

Išanalizavus SAS sistemos klasifikavimo su mokytoju procedūras ir makrokomandas bei nustačius jų trūkumus, išplėstos SAS sistemos galimybės programiniu įrankiu, leidžiančiu greičiau ir patogiau pagal klasifikavimo kokybę lyginti parametrinius tiesinės ir kvadratinės diskriminantinės analizės, neparametrinius branduolio ir artimiausių kaimynų bei logistinės regresinės analizės metodus.

Sukurtų programinių priemonių testavimas atliktas su skirtingais duomenimis. Siekiant išsiaiškinti parametrinių ir neparametrinių klasifikavimo su mokytoju metodų privalumus ir trūkumus bei taikymo galimybes, naudotos imtys, kurios skiriasi imties dydžiu, klasių atskiriamumu,

parametrinių klasifikavimo su mokytoju metodų taikymo prielaidų tenkinimu. Įrankio testavimas parodė, kad jis sprendžia darbe suformuluotas užduotis. Kiekvienu atveju buvo nustatyti tinkamiausi klasifikavimo metodai ir įvertinta jų klasifikavimo kokybė.

Darbe pasiūlyta metodika bei ją įgyvendinančios programinės priemonės patogios klasifikavimo su mokytoju metodų klasifikavimo kokybei lyginti. Pakanka nurodyti imtį, priklausomą bei diskriminavimo kintamuosius ir skirtingais metodais bus atliktas klasifikavimas bei įvertinta metodų klasifikavimo kokybė. Sukurtą įrankį galima lengvai papildyti naujais klasifikavimo su mokytoju metodais (pvz., klasifikavimo medžiais ar neuroniniais tinklais) ir naujais klaidingo klasifikavimo tikimybės vertinimo metodais.

1 BENDROJI DALIS

1.1 Klasifikavimo su mokytoju metodų apžvalga

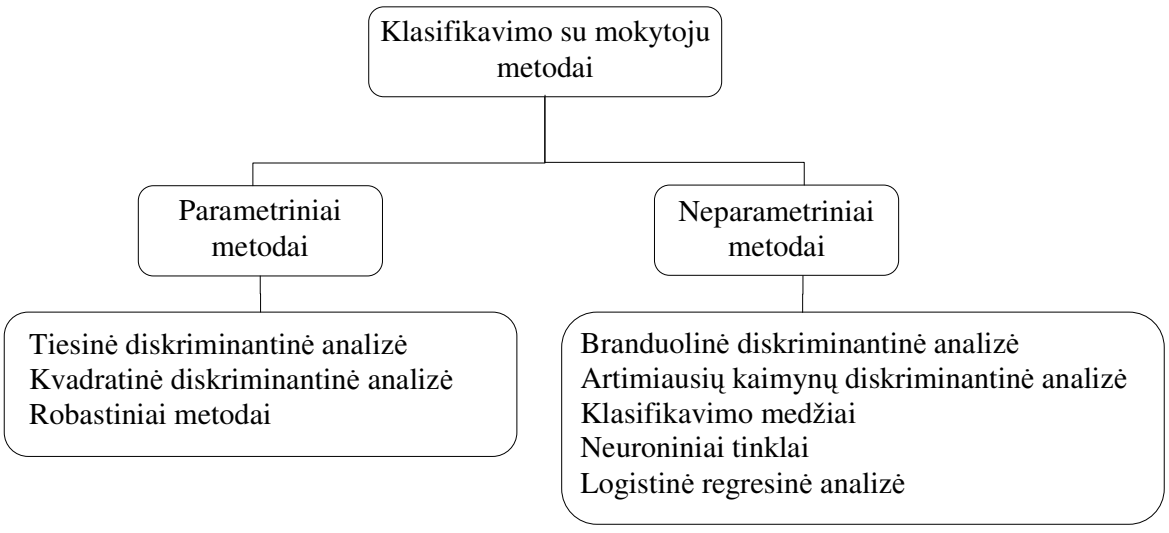
Klasifikavimo metodai skirstomi į dvi klases: klasifikavimas be mokytojo (angl. *unsupervised classification*) ir klasifikavimas su mokytoju (angl. *supervised classification*) [18, 21]. Darbe nagrinėjami tik klasifikavimo su mokytoju metodai, todėl toliau vartodami terminą klasifikavimas turėsime omeny klasifikavimą su mokytoju.

Literatūroje išskiriami trys pagrindiniai klasifikavimo su mokytoju etapai:

- diskriminavimo kintamųjų parinkimas,
- klasifikavimo taisyklių sudarymas,
- klasifikavimo kokybės įvertinimas.

Atliekant klasifikavimą su mokytoju pirmiausia reikia parinkti požymius, diskriminuojančius (atskiriančius) tiriamų objektų klases. Iš daugybės požymių parenkami tie, kurie geriausiai diskriminuoja tiriamų objektų klases. Literatūroje pateikta daug įvairių kriterijų pagal kuriuos sprendžiama apie požymių diskriminavimo savybes. Naudojama vienfaktorė dispersinė analizė bei Vilkso statistika [7]. Taip pat pasiūlyta nauja metodika, kurioje naudojama MiPP statistika (angl. *the misclassification-penalized posterior*), t.y. aposteriorinių tikimybių įverčių sumos ir klaidingai klasifikuotų objektų skirtumas [30]. Šiame darbe požymių parinkimas nenagrinėjamas.

Pagrindinis klasifikavimo etapas yra klasifikavimo taisyklių sudarymas. Darbe analizuojami dažniausiai praktikoje naudojami parametriniai bei neparametriniai metodai.



1 pav. Pagrindiniai klasifikavimo su mokytoju metodai

Parametriniuose metoduose daroma prielaida apie diskriminavimo kintamųjų tankio funkcijos tipą kiekvienoje klasėje. Tuomet reikia įvertinti tik nežinomus tankių parametrus [14, 33]. Dažniausiai daroma prielaida, kad diskriminavimo kintamųjų skirstinys yra daugiamatis normalusis. Tačiau taip pat naudojamos prielaidos, kad diskriminavimo kintamieji pasiskirstę pagal eksponentinį ar kitą dėsnį [21].

Neparametriniuose metoduose prielaidos apie diskriminavimo kintamųjų skirstinius nedaromos. Taikant šiuos metodus, tankio funkcijos įvertinamos naudojant neparametrinius branduolio, artimiausių kaimynų ir kitus metodus [10, 11, 21]. Yra neparametrinių metodų (klasifikavimo medžiai, neuroniniai tinklai), kuriuose sudarant klasifikavimo taisykles nenaudojama informacija apie diskriminavimo kintamųjų skirstinius ir nevertinamos tankio funkcijos [12, 33].

Literatūroje klasifikavimo metodų ir įvairių jų modifikacijų yra pateikta labai daug (keli šimtai). Dėl darbo apimties šiame darbe apžvelgiami tik dažniausiai praktikoje naudojami. Retai naudojami metodai – Euklido atstumo klasifikatoriai (angl. *Euclidean distance classifier*), skirstinių mišiniams skirti metodai (angl. *methods based on mixtures densities*), polinominių ir potencialinių funkcijų klasifikatoriai (angl. *polynomial and potential function classifiers*) [25], atraminių vektorių metodai (angl. *Support vector machine*) [18, 25, 32] ir kiti šiame darbe nenagrinėjami.

Norint pasirinkti tinkamiausią klasifikavimo su mokytoju metodą reikia įvertinti klasifikavimo metodo kokybę. Tam naudojami klaidingo klasifikavimo tikimybės įverčiai. Kai yra tik dvi klasės, skaičiuojamas jautrumas bei specifiškumas ir braižomos ROC kreivės.

Tarkime turime atrinktus, geriausiai tiriamų objektų klases diskriminuojančius p požymius $X = (X^{(1)},..., X^{(p)})$. Požymius žymintys kintamieji yra intervaliniai ir vadinami nepriklausomais arba diskriminavimo kintamaisiais.

Klasifikavimo su mokytoju metodai iš kitų klasifikavimo metodų išsiskiria tuo, kad juos galima taikyti tik tada, kai iš anksto žinomas klasių skaičius l $(l \geq 2)$ ir yra n dydžio mokomoji imtis[10, 21]:

$$S = \begin{pmatrix} x_1^{(1)} & x_1^{(2)} & ... & x_1^{(p)} & y_1 \\ x_2^{(1)} & x_2^{(2)} & ... & x_2^{(p)} & y_2 \\ ... & ... & ... & ... & ... \\ x_n^{(1)} & x_n^{(2)} & ... & x_n^{(p)} & y_n \end{pmatrix},$$

čia $y_i = k \Leftrightarrow \left(x_i^{(1)}, x_i^{(2)},..., x_i^{(p)}\right) \in \Pi_k, (i = 1,...,n; \quad k = 1,...,l)$, $n = \sum_{k=1}^{l} n_k$, n_k – objektų iš k-tosios klasės skaičius, $\Pi_1,...,\Pi_l$ – nepersikertantys diskriminavimo kintamųjų erdvės poerdviai. Toliau darbe $\Pi_1,...,\Pi_l$ vadinsime objektų klasėmis, o Y – priklausomu kintamuoju (y_i yra jo realizacijos).

Pagrindinis klasifikavimo metodų tikslas – sudaryti klasifikavimo taisykles: $\delta : x \to \Pi_k, (k = 1,...,l)$, t.y. taisykles, kurios klasifikuojamus objektus pagal jų požymius priskirtų konkrečiai klasei (objektas vienu metu negali priklausyti kelioms klasėms). Objektas priskiriamas tai klasei, kurios didžiausia aposteriorinė tikimybė $\tau_i(x) = P(X \in \Pi_i \mid X = x)$, dažniausiai randama naudojant Bejeso formulę [10, 21]:

$$\tau_i(x) = \frac{\pi_i f_i(x)}{\sum_{i=1}^{l} \pi_i f_i(x)} \ (i=1,...,l). \qquad \textbf{(1.1)}$$

čia $\pi_i = P(X \in \Pi_i)$ – apriorinė tikimybė (tikimybė, kad atsitiktinai iš populiacijos parinktas objektas priklauso klasei Π_i), $f_1,...,f_l$ – *p*-matės *X* tankio funkcijos klasėse $\Pi_1,...,\Pi_l$. Akivaizdu, kad $\sum_{i=1}^{l} \pi_i = 1$.

Praktikoje nežinome nei π_i, nei f_i, todėl naudodami mokomąją imtį turime surasti jų įverčius $\hat{\pi}_i$ ir $\hat{f}_i$. Įstatę juos į Bejeso formulę 1.1 gauname aposteriorinės tikimybės įvertį $\hat{\tau}_i$, pagal kurį sprendžiame kuriai klasei objektas priklauso. Yra pasiūlyta keletas įvairių apriorinės tikimybės įverčių [14, 21, 28], tačiau dažniausiai naudojamas:

$$\hat{\pi}_i = n_i / n \ (i=1,...,l), \qquad \textbf{(1.2)}$$

čia n_i – objektų skaičius *i*-tojoje klasėje, *n* – bendras objektų skaičius.

1.1.1 Tiesinė diskriminantinė analizė

XX amžiaus ketvirtame dešimtmetyje R. Fišeris sukūrė pirmąjį parametrinį klasifikavimo metodą – tiesinę diskriminantinę analizę [21]. Taikant šį metodą, ieškoma tiesinė transformacija, kuri maksimizuotų sklaidą tarp klasių ir minimizuotų sklaidą klasių viduje. Tai vienas paprasčiausių ir lengviausiai interpretuojamų klasifikavimo metodų. Tačiau norint jį taikyti, turi būti tenkinamos prielaidos [28]:

1. $f_i(x) = \frac{1}{(2\pi)^{p/2} \mid \sum \mid^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \sum^{-1} (x-\mu_i) \right)$ *(i=1,...,l),*
2. $\sum_i = \sum$ *(i=1,...,l),*

čia $\sum$ - jungtinė kovariacijų matrica (angl. *the pooled covariance matrix*), $\sum_i$ – kovariacijų matrica *i*-tojoje klasėje, μ_i - *X* vidurkis *i*-toje klasėje.

Iš Bejeso formulės 1.1 seka, kad objektas priskiriamas klasei Π_k, kuriai $\pi_k f_k(x) \geq \pi_i f_i(x)$ (*i=1,…,l*). Tuo pasinaudojant sudaroma diskriminavimo taisyklė, priskirianti objektą klasei Π_k, kurios dydis $d_k^L(x)$ yra maksimalus. Čia

$$d_i^L(x) = \mu_i^T \Sigma^{-1} x - \frac{1}{2}\mu_i^T \Sigma^{-1} \mu_i + \ln(\pi_i) \quad \textbf{(1.3)}$$

yra tiesinė funkcija nuo *x*. Todėl šis metodas vadinamas tiesine diskriminantine analize.

Panaudojant mokomąją imtį ieškomi parametrų Σ ir μ_i įverčiai $\hat{\Sigma}$ ir $\hat{\mu}_i$. Dažniausiai naudojami [21, 28]:

$$\hat{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{i_j} \quad (i=1,...,l), \quad \textbf{(1.4)}$$

$$\hat{\Sigma} = \frac{\sum_{i=1}^{l}(n_i - 1)\hat{\Sigma}_i}{\sum_{i=1}^{l} n_i - l}, \quad \textbf{(1.5)}$$

$$\hat{\Sigma}_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{i_j} - \hat{\mu}_i)(x_{i_j} - \hat{\mu}_i)^T \quad (i=1,...,l), \quad \textbf{(1.6)}$$

čia x_{i_j} – *i*-tosios klasės *j*-tasis objektas. Įstačius parametrų įverčius į formulę 1.3, gaunamos klasifikavimo taisyklės.

1.1.2 Kvadratinė diskriminantinė analizė

Parametrinis kvadratinės diskriminantinės analizės metodas nuo tiesinės skiriasi tik tuo, kad nereikalaujama, jog skirtingų klasių kovariacijų matricos būtų identiškos [10, 14, 21]. Taikant šį metodą daroma prielaida $f_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i)\right)$ *(i=1,...,l).* Klasifikavimo taisyklės sudaromos naudojant kvadratinę funkciją:

$$d_i^Q(x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma^{-1}(x-\mu_i) + \ln(\pi_i). \quad \textbf{(1.7)}$$

Objektas priskiriamas klasei Π_k, kurios $d_k^Q(x)$ yra maksimalus. Parametrai μ_i ir Σ_i dažniausiai įvertinami pagal 1.4 ir 1.6 formules.

1.1.3 Robastiniai metodai

Tiek tiesinėje, tiek ir kvadratinėje diskriminantinėje analizėje pagal 1.4-1.6 formules surasti įverčiai $\hat{\Sigma}_i$ ir $\hat{\mu}_i$ yra jautrūs duomenų išskirtims. Siekiant išvengti neigiamo išskirčių poveikio, naudojami robastiniai metodai, t.y. metodai mažai jautrūs išskirtims. Pavyzdžiui, viename robastiniame metode parametrų Σ_i ir μ_i įverčiai apskaičiuojami pagal 1.4-1.6 formules, tačiau naudojami ne visi klasių objektai, o tik jų dalis. Kiekvienoje klasėje atmetama $[\alpha \cdot n_i]$ daugiausiai nuo klasės centro nutolusių ir $[\alpha \cdot n_i]$ mažiausiai nutolusių objektų (naudojamas Euklido arba Mahalanobio atstumo matas). Jei turime stebėjimų $\{x_j\}_{j=1}^{n_i}$ variacinę eilutę $x_{(1)} \le x_{(2)} \le ... \le x_{(n_i)}$, tai $\hat{\mu}_\alpha = (n - 2[\alpha \cdot n_i])^{-1} \sum_{j=[\alpha \cdot n_i]}^{n_i - [\alpha \cdot n_i]} x_{(j)}, 0 < \alpha < 1/2$ [11]. M. Hubert ir K. V. Driessen panaudodami šią metodiką sukūrė dar efektyvesnį robastinės kvadratinės diskriminantinės analizės metodą [14]. Plačiau robastiniai klasifikavimo su mokytoju metodai pateikiami literatūroje [25].

1.1.4 Branduolinė diskriminantinė analizė

Taikant neparametrinį branduolinės diskriminantinės analizė metodą, naudojama Bejeso formulė 1.1, kurioje nežinomos tankio funkcijos $f_i(x)$ yra įvertinamos branduolio metodu. *P*-matės tankio funkcijos $f_i(x)$ branduolio įvertis taške *x* [11, 28]:

$$\hat{f}_i^{(K_p)}(x) = \frac{1}{n_i |H_i|} \sum_{j=1}^{n_i} K_p\left(\frac{x - x_{i_j}}{H_i}\right), \qquad \textbf{(1.8)}$$

čia x_{i_j} – *i*-tosios klasės *j*-tasis objektas, K_p – *p*-matė branduolio funkcija, kuri suintegruojama į vienetą, H_i – nesinguliari *i*-tosios klasės glodinimo parametrų matrica, $H_i \in M_{p \times p}$. Dažniausiai parenkama neneigiama, tolydi ir simetrinė branduolio funkcija. Taigi tankio funkcijos $f_i(x)$ branduolio įvertis taške *x* yra svertinis objektų iš $S \cap \Pi_i$, pakliūnančių į lokalią *x* aplinką, vidurkis. Lokali aplinka apibrėžiama H_i pagalba, o svoriai objektams suteikiami branduolio funkcijos pagalba.

Kadangi su daugiamate branduolio funkcija sudėtinga dirbti, tai dažnai daroma prielaida: $K_p(x) = \prod_{i=1}^{p} h_i^{-1} K_1\left(\frac{x^{(i)}}{h_i}\right)$, čia $x^{(i)}$ – *i*-toji vektoriaus *x* koordinatė, K_1 – vienmatė branduolio funkcija.

Klasėse naudojant vieną glodinimo parametrą h_i ($H_i = h_i I_p$), branduolio įvertis:

$$\hat{f}_i^{(K_1)}(x) = \frac{1}{n_i h_i^p} \sum_{j=1}^{n_i} \prod_{k=1}^{p} K_1\left(\frac{x^{(k)} - x_{i_j}^{(k)}}{h_i} \right). \quad \textbf{(1.9)}$$

Šiuo atveju visos vektoriaus x koordinatės turėtų būti panašaus dydžio, nes priešingu atveju susidurtume su nevienodos skirtingai matuojamų požymių įtakos problema [3].

Taikant branduolio metodą iškyla dvi problemos: branduolio funkcijos ir optimalaus glodinimo parametro parinkimas. Jų sprendimo būdai pateikti literatūros šaltiniuose [11, 21, 28].

1.1.5 Artimiausių kaimynų diskriminantinė analizė

Vienas paprasčiausių neparametrinės diskriminantinės analizės metodų yra k artimiausių kaimynų metodas. Panagrinėkime paprasčiausią artimiausio kaimyno metodą (kai $k=1$). Tarkime turime mokomąją imtį $(x_1, y_1),...,(x_n, y_n)$. Tegu $d_i = \| x - x_i \|$ atstumas tarp x ir x_i. Surandame tokį x_i', kad d_i būtų mažiausias ($i = 1,2,...,n$), t.y. randame artimiausią kaimyną. Tada objektą priskiriame tai pačiai klasei kaip ir x_i'. Kai $k \neq 1$, randame k artimiausių kaimynų ir klasifikuojamą objektą priskiriame tai klasei, kurios dažnis didžiausias [6]. Šiuo atveju iškyla optimalaus parametro k parinkimo problema [11, 21]. Jeigu parenkame labai mažą, tai atsižvelgiame tik į dalį turimos informacijos, t.y. į dalį mokomosios imties. Jeigu parenkame k artimą mokomosios imties dydžiui, tai klasifikuojamus objektus visada priskiriame tai klasei, kurios dažnis imtyje didžiausias. Pagrindinis metodo trūkumas yra tas, kad sudėtinga jį taikyti, kai turime didelę mokomąją imtį, nes reikia atlikti daug skaičiavimų.

1.1.6 Klasifikavimo medžiai

Klasifikavimo medžiai – tai grupė vienų lanksčiausių ir labiausiai paplitusių neparametrinių klasifikavimo metodų. Dėl paprastumo ir lengvo interpretavimo būtent juos renkasi daugelis kompanijų, susiduriančių su klasifikavimo bei sprendimų priėmimo problemomis. Vienas jų pranašumų lyginant su kitais klasifikavimo metodais yra tas, kad diskriminavimo kintamieji gali būti ne tik intervalų, bet ir vardų skalėje [39].

Pagrindiniai metodai priskiriami šiai grupei yra CHAID (angl. *Chi-squared automatic interaction Detector*), CART (angl. *classification and regression trees*), QUEST (angl. *quick ubiased efficient tree*) [1]. Taip pat yra įvairių jų modifikacijų [21, 32, 38]. Visuose šiuose metoduose yra naudojami sprendimų priėmimo medžiai.

Sprendimų priėmimo medis – tai diagrama (ciklų neturintis orientuotas grafas), kuri iliustruoja sprendimų priėmimo taisykles. Ji prasideda viršūne, vadinama šaknimi (angl. *root node*), kurioje yra visa turima mokomoji imtis *S*. Iš šaknies išeina šakos, jungiančios šaknį su dukterinėmis viršūnėmis, kuriose yra imtys *t* (mokomosios imties subimtys $t \subseteq S$). Viršūnės iš kurių šakos neišeina vadinamos lapais (angl. *leaves, terminal nodes*). Kiekvienos viršūnės (išskyrus lapų) imties padalinimui naudojamas padalinimo funkcija (angl. *split*) ir padalinimo kintamasis (vienas iš diskriminavimo kintamųjų). Atsižvelgiant į padalinimo funkcijas ir kintamuosius, leidžiamasi iš šaknies žemyn kol pasiekiamas lapas, priklausomai nuo kurio yra priimamas vienoks ar kitoks sprendimas.

Sprendimų priėmimo medžiai skirstomi į binarinius ir nebinarinius. Pirmuose iš kiekvienos viršūnės, išskyrus lapus, išeina dvi šakos, o antruose dvi arba daugiau nei dvi. Binariniame medyje, kai padalinimo kintamasis *x* intervalinis, padalinimo funkcija *s* yra dvireikšmė:

$$s(x) = \begin{cases} 1, & jei\ x \le c \\ 0, & jei\ x > c \end{cases}, \forall x \in t,$$

čia *c*=*const.* Kai padalinimo kintamasis kategorinis ir gali įgyti $\{a_1,...,a_d\}$ kategorijas, tai *s* yra:

$$s(x) = \begin{cases} 1, & jei\ x \in M \\ 0, & jei\ x \notin M \end{cases}, \forall x \in t,$$

čia $M \subseteq \{a_1,...,a_d\},$. Naudojant padalinimo funkciją *s* ir padalinimo kintamąjį *x*, imtis $t \subseteq S$ dalina į dvi subimtis (t_t ir t_f) taip:

$$\forall x \in t : s(x) = \begin{cases} 1 \Rightarrow x \in t_t \\ 0 \Rightarrow x \in t_n \end{cases},$$

čia $t_t \cap t_n = \varnothing, t_t \cup t_n = t$.

CHAID yra vienas pirmųjų klasifikavimo medžių metodų [32]. Pagrindinis šio metodo privalumas, kad jame naudojamas nebinarinis medis. Dėl to sukuriamas platesnis sprendimų priėmimo medis nei kitais metodais. Metodo trūkumas tas, kad, skirtingai nei kituose dviejuose metoduose, tas pats diskriminavimo kintamasis imtims dalinti gali būti panaudotas tik vieną kartą. Todėl dažnai gaunamas per mažas medis. Šis metodas taip pat pasižymi didžiausia padalinimo kintamųjų parinkimo paklaida, t.y. kai padalinimo kintamaisiais parenkami daugiausia kategorijų turintys diskriminavimo kintamieji. CHAID algoritmą sudaro trys pagrindiniai žingsniai [1]:

1. *Kintamųjų kategorizavimas*. Visi intervaliniai diskriminavimo kintamieji transformuojami į kategorinius.
2. *Kategorijų sujungimas*. Naudojant χ^2 homogeniškumo kriterijų, sujungiamos kiekvieno diskriminavimo kintamojo kategorijos, kurios statistiškai reikšmingai nesiskiria priklausomojo kintamojo atžvilgiu.

3. *Padalinimo kintamojo parinkimas*. Naudojant Bonferonio nepriklausomumo kriterijų randamas diskriminavimo kintamasis, kurio *p*-reikšmė yra mažiausia. Jei $p < \alpha_s$ (α_s - iš anksto apibrėžtas padalinimo slenkstis), tai imtis dalinama į subimtis pagal surasto diskriminavimo kintamojo kategorijas. Jei $p > \alpha_s$, tai viršūnė yra lapas.

CART šiuo metu yra labiausiai paplitęs klasifikavimo medžių metodas. Pagrindinis jo privalumas yra tas, kad jis išanalizuoja visus galimus imčių padalinimus ir išrenka geriausią. Taigi jis optimaliai klasifikuoja mokomąją imtį (tačiau tai dar nereiškia, kad toks klasifikavimas yra optimalus populiacijoje). Žvelgiant iš kitos pusės, tai tampa vienu iš trūkumų, nes tokia optimalaus padalinimo paieška užtrunka ilgai. Be to gaunamas labai didelis binarinis sprendimų priėmimo medis [1].

Metode išskiriami du pagrindiniai etapai: medžio užauginimo ir apgenėjimo. Pirmajame, pradedant šaknimi, kiekviena viršūnė dalinama į dukterines. Čia tenka spręsti padalinimo funkcijos ir kintamojo parinkimo problemą. Ieškant optimalaus padalinimo nagrinėjami visi kintamieji bei visi galimi jų padalinimai. Optimali padalinimo funkcija parenkama maksimizuojant Gini indeksą, Entropijos arba kitą funkciją Φ [1, 8]. Tokiu būdu medis auginamas tol, kol lapuose lieka tik vienos klasės objektai arba kol $\max_s \Phi(s,t) \geq \beta$, čia $\beta \geq 0$ – slenkstinė reikšmė. Etapo pabaigoje gauname maksimalų sprendimų priėmimo medį T_{max}. Tačiau toks medis būna labai didelis ir netinkamas taikymams. Jis labai įtakojamas paklaidų imtyje. Todėl naudojamas taip vadinamas medžio apgenėjimas.

Medžio apgenėjimo pagrindinis tikslas – kiek galima sumažinti sprendimų priėmimo medį, neprarandant diskriminavimo savybių, t.y. žymiai nepadidinant klaidingo klasifikavimo tikimybės. Tam naudojamas sudėtingumo matas (angl. *cost-complexity*) [1]:

$$R_\alpha(T) = R(T) + \alpha \cdot \left|\tilde{T}\right|, \qquad \textbf{(1.10)}$$

čia *R(T)* –klaidingo klasifikavimo tikimybė, kai klasifikavimui naudojamas medis *T*, $\tilde{T}$ – lapų aibė, α – medžio sudėtingumo parametras. Taigi sudėtingumo matas įvertina klaidingo klasifikavimo riziką bei medžio sudėtingumą. Didėjant medžio sudėtingumui mažėja klaidingo klasifikavimo rizika ir atvirkščiai. Reikia rasti optimalų variantą.

Kiekvienam α egzistuoja optimalus medis sudėtingumo mato prasme: $T(\alpha) = \arg\min_{T \subseteq T_{\max}} R_\alpha(T)$. Kai $\alpha \approx 0$, tai $T(\alpha) = T_{\max}$. Toliau didinant α ir apskaičiuojant optimalius medžius, gauname seką $T_{\max} = T_0 \supseteq T_1 \supseteq ... \supseteq \{t_0\}$. Iš sekos parenkamas medis, kurio Bejeso rizikos indekso 1.24 įvertis yra mažiausias.

Naujausias iš klasifikavimo medžių metodų yra QUEST [1, 32]. Jis turi dauguma CART metodo privalumų, tačiau tuo pačiu ir vieną pagrindinių trūkumų – juo sukurtas sprendimų priėmimo medis labai didelis. Pagrindinis jo privalumas – greitis. Esant dideliam skaičiui nepriklausomų kintamųjų su

daug kategorijų QUEST sprendimų priėmimo medis sudaromas net iki kelių šimtų kartų greičiau nei CART. Be to šis metodas turi mažiausią padalinimo kintamųjų parinkimo paklaidą.

1.1.7 Neuroniniai tinklai

Dirbtiniai neuroniniai tinklai yra supaprastinti centrinės nervų sistemos modeliai. Tai tinklai, sudaryti iš tampriai tarpusavyje susietų elementų (neuronų), galinčių reaguoti į įeinantį signalą ir išmokti derintis prie aplinkos. Labiausiai paplitusios yra dvi dirbtinių neuroninių tinklų rūšys [18, 21, 25].

- Vienasluoksnis perceptronas (angl. *SLP single layer perceptron*) – tiesiog vienas neuronas.
- Daugiasluoksnis perceptronas (angl. *MLP multilayer perceptron*) – daug neuronų išdėstytų sluoksniais. Kiekvieno sluoksnio neuronų išėjimai sujungti su kito sluoksnio neuronų įėjimais. Įėjimo sluoksnis – pradiniai duomenys (diskriminavimo kintamųjų reikšmės), išėjimo sluoksnis – paskutiniame sluoksnyje esančių neuronų išėjimai (priklausomojo kintamojo reikšmės); visi kiti sluoksniai vadinami paslėptaisiais.

Perceptronas (neuronas) susideda iš įėjimo (diskriminavimo) kintamųjų, juos atitinkančių svorių w_i ($i=0,1,\ldots,p$) ir aktyvacijos funkcijos f_a. Įėjimo kintamųjų ir atitinkamų svorių sandaugų sumą paveikus aktyvacijos funkcija, gaunamas išėjimo kintamasis:

$$Y = f_a\left(w_1 X^{(1)} + \ldots + w_p X^{(p)} + w_0\right). \qquad \textbf{(1.11)}$$

Klasifikuojant dažniausiai naudojama sigmoidinė aktyvacijos funkcija: $f_a(x) = 1/(1+e^{-x})$ [21, 25]. Tokiu atveju išėjimo kintamasis $Y \in [0;1]$. Todėl naudojant vienasluoksnį perceptroną, galima spręsti klasifikavimo uždavinius, kai $l=2$. Objektas priskiriamas pirmajai klasei jei $y < 0.5$, priešingu atveju – antrajai.

Kai klasių yra daugiau nei dvi, reiktų naudoti daugiasluoksnius perceptronus, paskutiniame sluoksnyje turinčius l neuronų. Pakanka perceptronų su vienu paslėptu sluoksniu, nes toks perceptronas yra universalus aproksimatorius. Juo galima gauti bet kokio sudėtingumo atskyrimo paviršių [25]. Tuo tarpu su vienasluoksniu perceptronu gauname tik tiesinius atskyrimo paviršius. Daugiasluoksnio perceptrono su vienu paslėptu sluoksniu r-tojo išėjimo dydžio Y_r reikšmė apskaičiuojama taip [18]:

$$Z_k = f_{a1}\left(\sum_{j=1}^{p} w_{jk}^{(z)} X^{(j)} + w_{0k}^{(z)}\right),\ k=(1,2,\ldots h), \qquad Y_r = f_{a2}\left(\sum_{i=1}^{h} w_{ir}^{(y)} Z_i + w_{0r}^{(y)}\right),\ r=(1,2,\ldots l), \qquad \textbf{(1.12)}$$

čia h – neuronų skaičius paslėptame sluoksnyje, f_{a1}, f_{a2} – aktyvacijos funkcijos. Jei f_2 sigmoidinė, tai objektas priskiriamas r-tajai klasei, kur $r = \arg\max_{i\in\{1,\ldots l\}} y_i$.

Klasifikavimo, taikant neuroninius tinklus, problema – perceptrono koeficientų *W* radimas. Tam naudojamas taip vadinamas perceptrono apmokymas, t.y. parenkami svoriai minimizuojantys nuostolių funkciją. Naudojamos įvairios nuostolių funkcijos [18, 21]. Vienasluoksnio perceptrono atveju dažnai naudojama nuostolių f-ja:

$$c = \sum_{j=1}^{n} \left(y_j - f_a(w_1 x_j^{(1)} + ... + w_p x_j^{(p)} + w_0) \right)^2 . \qquad \textbf{(1.13)}$$

Daugiasluoksnių perceptronų nuostolių funkcijos yra daug sudėtingesnės. Nuostolių funkcijų minimizavimui yra naudojami įvairūs metodai – jungtinių gradientų (angl. *conjugate gradient*), atbulinės sklaidos (angl. *error back propagation*) ir kiti [25].

Atbulinės sklaidos (AS) metodas pagrįstas gradientiniu nusileidimu, kuris modifikuoja svorius taip, kad būtų sumažinta sisteminė paklaida. Pirmiausia parenkami pradiniai svoriai. Literatūros šaltiniuose [18, 25] rekomenduojamos mažos atsitiktinai sugeneruotos reikšmės. Tada kiekvienas mokomosios imties objektas įvedamas į tinklą ir sluoksnis po sluoksnio suskaičiuojama išėjimo reikšmė, kuri lyginama su turima ar norima reikšme bei skaičiuojama paklaida (nuostolių f-jos reikšmė). Gautos paklaidos naudojamos kaip įėjimo kintamieji išėjimo sluoksnio jungtims, nuo kurių pradedama keisti svorius sluoksnis po sluoksnio atbuline kryptimi.

Naudojant AS metodą, paslėptųjų sluoksnių svoriai modifikuojami naudojant sekančio sluoksnio paklaidas. Todėl paklaidos, suskaičiuotos išėjimo sluoksnyje, naudojamos svoriams tarp paskutinio paslėptojo ir išėjimo sluoksnių pakeisti. Analogiškai paskutinio paslėptojo sluoksnio paklaidos naudojamos prieš tai esančio sluoksnio svoriams pakeisti ir taip toliau, kol pakoreguojami pirmojo paslėptojo sluoksnio svoriai. Tokiu būdu paklaidos yra skleidžiamos atgal sluoksnis po sluoksnio nuosekliai darant pakeitimus atitinkamiems svoriams. Šis procesas kartojamas kol nepasiekiamas pasirinktas sustojimo kriterijus. Literatūros šaltiniuose [18, 21, 25] rekomenduojama procesą kartoti:

- fiksuotą iteracijų skaičių;
- kol paklaida nesumažėja iki norimo lygio;
- kol testinėje imtyje paklaida nepradeda didėti.

Naudojant paskutinį sustojimo kriterijų, perceptronas apmokomas su apmokymo imties objektais, o testuojamas (tikrinamas paklaidos dydis) su testinės imties objektais. Apie tetinės ir apmokymo imties sudarymą žiūrėti skyrelyje 1.10.

Klasifikavimo su mokytoju uždavinių sprendimui be minėtųjų vienasluoksnių ir daugiasluoksnių perceptronų naudojama daugybė įvairių neuroninių tinklų modelių bei jų modifikacijų: bazinių radialinių funkcijų tinklai (angl. *radial basis function networks*) [25], tikimybiniai neuroniniai tinklai (angl. *probabilistic neural networks*) [4], mokomųjų vektorių kvantavimo tinklai (angl. *learning vector quantisation networks*) ir kiti [21].

1.1.8 Logistinė regresija

Kai netenkinama normalumo sąlyga, tai viena iš alternatyvų klasikinei tiesinei diskriminantinei analizei yra logistinė regresija. Ji skirta kategorinio kintamojo reikšmių tikimybių prognozavimui. Pagal priklausomojo kintamojo įgyjamų reikšmių aibę logistinė regresija dažniausiai skirstoma į klases:

- dichotominė, kai priklausomas kintamasis gali įgyti tik dvi reikšmes (angl. *binary responses*);
- politominė rangų, kai priklausomo kintamojo reikšmės yra rangų skalėje ir jų daugiau nei dvi (angl. *ordinal responses*);
- politominė vardų, kai priklausomo kintamojo reikšmės yra vardų skalėje ir jų daugiau nei dvi (angl. *nominal responses*).

Pagal nepriklausomų kintamųjų skaičių logistinė regresija skirstoma į vienmatę (kai yra vienas nepriklausomas (diskriminavimo) kintamasis) ir daugiamatę (kai yra daugiau nei vienas nepriklausomas kintamasis).

Tarkime, priklausomas kintamasis Y gali įgyti tik 1 arba 0 su atitinkamomis tikimybėmis τ_1 ir $1-\tau_1$. Tada dichotominės logistinės regresijos modelis yra:

$$\tau_1 = \frac{e^{\beta_0+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)}}}{1+e^{\beta_0+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)}}}, \qquad \textbf{(1.14)}$$

$$\ln\frac{\tau_1}{1-\tau_1} = \beta_0+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)}. \qquad \textbf{(1.15)}$$

Santykis $\frac{\tau_1}{1-\tau_1}$ vadinamas galimybe (galimybės įvertinimu) įvykti įvykiui $Y=1$. Įvykio $Y=1$ galimybė įvykti yra didesnė už 1, kai $P(Y=1) > P(Y=0)$ [7].

Politominės rangų logistinės regresijos matematinis modelis, kai priklausomas kintamasis gali įgyti l reikšmių [24]:

$$\begin{aligned}
&\ln\frac{\tau_1}{1-\tau_1} = \beta_{01}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)},\\
&\ln\frac{\tau_1+\tau_2}{1-\tau_1-\tau_2} = \beta_{02}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)},\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\ln\frac{\tau_1+\tau_2+\ldots+\tau_{l-1}}{1-\tau_1-\tau_2-\ldots-\tau_{l-1}} = \beta_{0l-1}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+\ldots+\beta_p X^{(p)},\\
&\tau_1+\tau_2+\ldots+\tau_{l-1}+\tau_l = 1.
\end{aligned} \qquad \textbf{(1.16)}$$

Nesunku pastebėti, kad galimybės $\frac{\tau_1+\tau_2+...+\tau_j}{1-\tau_1-\tau_2-...-\tau_j}, j=1,2,...,l-1$ yra proporcingos, t.y. $\frac{\tau_1+\tau_2+...+\tau_j}{1-\tau_1-\tau_2-...-\tau_j} = const \cdot \frac{\tau_1+\tau_2+...+\tau_{j-1}}{1-\tau_1-\tau_2-...-\tau_{j-1}}, j=2,3,...,l-1$. Todėl šis modelis dažnai vadinamas lygiagrečiu regresiniu modeliu (angl. *parallel regression model*) arba proporcingų galimybių modeliu (angl. *proportional odds model*) [24]. Iš 1.16 formulių gauname tikimybių išraiškas:

$$\tau_1 = \frac{e^{\beta_{01}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}}{1+e^{\beta_{01}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}},$$

$$\tau_1+\tau_2 = \frac{e^{\beta_{02}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}}{1+e^{\beta_{02}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}}, \qquad \textbf{(1.17)}$$

……..

$$\tau_1+\tau_2+...+\tau_{l-1} = \frac{e^{\beta_{0l-1}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}}{1+e^{\beta_{0l-1}+\beta_1 X^{(1)}+\beta_2 X^{(2)}+...+\beta_p X^{(p)}}},$$

$$\tau_l = 1-(\tau_1+\tau_2+...+\tau_{l-1}).$$

Politominės vardų logistinės regresijos matematinis modelis, kai priklausomas kintamasis gali įgyti *l* reikšmių [9]:

$$\ln\frac{\tau_i}{\tau_l} = \beta_{0i}+\beta_{1i}X^{(1)}+\beta_{2i}X^{(2)}+...+\beta_{pi}X^{(p)}, i=1,2,...,l-1, \qquad \textbf{(1.18)}$$

$$\tau_1+\tau_2+...+\tau_{l-1}+\tau_l = 1.$$

Atitinkamos tikimybių išraiškos:

$$\tau_i = \frac{e^{u_i}}{1+e^{u_1}+...+e^{u_m}},\ u_i = \beta_{0i}+\beta_{1i}X^{(1)}+\beta_{2i}X^{(2)}+...+\beta_{pi}X^{(p)}, \qquad \textbf{(1.19)}$$

$$i=1,2,...,l-1,\ \tau_l = 1-(\tau_1+\tau_2+...+\tau_{l-1}).$$

Visais logistinės regresijos atvejais parametrų $\beta=(\beta_0,\beta_1,\beta_2,...,\beta_p)$ įverčiai $\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,...,\hat{\beta}_p)$ randami panaudojant didžiausio tikėtinumo metodą. Tai iteracinis procesas, kai pasirenkamos pradinės įverčių reikšmės ir taikant tam tikrą skaičiavimo metodą keičiamos tol, kol stabilizuojasi [2, 7].

Naudojant Voldo kriterijų, galime patikrinti statistinę hipotezę $\begin{cases}H_0:\beta_i=0\\H_1:\beta_i\neq 0\end{cases}$, o panaudojant χ^2 kriterijų – hipotezę $\begin{cases}H_0:\beta_1=\beta_2=...=\beta_p=0\\H_1:bent\quad vienas\quad \beta_i\neq 0, i\in\{1,2,...,p\}\end{cases}$ [2, 7].

Pagrindiniai koeficientai, nusakantys logistinės regresijos suderinamumą su duomenimis:

Makfadeno pseudodeterminacijos

$$r_M^2 = 1 - \frac{\ln L(\hat{\beta})}{\ln L(\hat{\beta}_0, 0)}, \quad \textbf{(1.20)}$$

Kokso-Snelo

$$r_M^2 = 1 - \left(\frac{\ln L(\hat{\beta}_0, 0)}{\ln L(\hat{\beta})} \right)^{2/n}, \quad \textbf{(1.21)}$$

čia $L(\hat{\beta})$ – didžiausio tikėtinumo funkcijos maksimumas, kai $\hat{\beta}$ yra didžiausio tikėtinumo metodu rasti koeficientų įverčiai, $L(\hat{\beta}_0, 0)$ – didžiausio tikėtinumo funkcijos maksimumas, kai pasirinktas logistinės regresijos modelis, kuriame $\beta_1 = \beta_2 = ... = \beta_p = 0$ [7].

Visus paminėtus logistinės regresijos modelius sieja prielaida apie įvykio galimybės logaritmo ir nepriklausomų kintamųjų tiesinę priklausomybę. Tačiau daugeliu atveju įvykio galimybės logaritmą ir nepriklausomus kintamuosius sieja netiesinė priklausomybė. Tokiu atveju naudojamas logistinės regresijos modelis [10, 24]:

$$\ln \frac{\tau_1}{1 - \tau_1} = g(x), \quad \textbf{(1.22)}$$

čia *g(x)* – nežinoma funkcija, kuri įvertinama branduolio, splainų ar kitu metodu.

1.1.9 Klaidingo klasifikavimo tikimybės ir jų įverčiai

Klasifikavimo taisyklėms keliamas reikalavimas minimizuoti klaidingo klasifikavimo tikimybę $\alpha_{ij} = P(\Pi_j \mid \Pi_i)$ – tai yra tikimybę *i*-tosios klasės objektą priskirti *j*-tajai klasei (naudojantis nustatyta klasifikavimo taisykle δ). Praktikoje dažnai pasitaiko, kad daug svarbiau yra minimizuoti α_{ij} nei α_{ji}, *i, j* – fiksuoti. Pvz., daugeliu atveju geriau sveikam žmogui diagnoziuoti ligą, nei sergančiam šios ligos nediagnozuoti. Tuo tikslu įvedamos vadinamosios klaidingo klasifikavimo kainos c_{ij}. Tada *i*-tosios klasės klaidingo klasifikavimo tikimybė [21]:

$$R_i(\delta) = \sum_{j=1, j \neq i}^{l} c_{ij} \alpha_{ij} \quad (i=1,...,l). \quad \textbf{(1.23)}$$

Susumavus visų klasių klaidingo klasifikavimo tikimybes, gauname bendrą klaidingo klasifikavimo tikimybę (Bejeso rizikos koeficientą):

$$R(\delta) = \sum_{i=1}^{l} \pi_i R_i(\delta). \quad \textbf{(1.24)}$$

Toliau darbe kalbėdami apie bendrą klaidingo klasifikavimo tikimybę žodį „bendra“ praleisime.

Paprasčiausiu atveju, kai yra tik dvi klasės Π_1, Π_2, vienas diskriminavimo kintamasis ir $c_{12} = c_{21} = 1$, tai klaidingo klasifikavimo tikimybė $R(\delta) = \pi_1 \int_{\Pi_2} f_1(X)dX + \pi_1 \int_{\Pi_1} f_2(X)dX$. Sprendžiant realius uždavinius, nežinome nei π_i, nei f_i, (*i=1,2).* Todėl panaudoję jų įverčius $\hat{\pi}_i$ ir $\hat{f}_i$, gauname klaidingo klasifikavimo tikikimybės įvertį $\hat{R}$. Šis įvertis priklauso nuo pasirinktos mokomosios imties, todėl vadinamas sąlygine klaidingo klasifikavimo tikimybe (angl. *conditional probability of misclassification*) arba bendra sąlygine paklaida (angl. *conditional generalisation error*) [25]. Toliau darbe $\hat{R}$ bus vadinama tiesiog klaidingo klasifikavimo tikimybės įverčiu.

Pagrindiniai klaidingo klasifikavimo tikimybės vertinimo metodai:

- savos imties (angl. *resubstition*),
- testinės imties,
- kryžminio patikrinimo (angl. *cross-validation*),
- Monte Karlo kryžminio patikrinimo (angl. *Monte Carl cross-validation*),
- įkelčių (angl. *bootstrap*).

Pirmasis ir paprasčiausias yra savos imties metodas. Pagrindinis šio metodo trūkumas – taikant šį metodą gaunami „optimistiški“ (paslinkti į mažesnę pusę) klaidingo klasifikavimo tikimybės įverčiai, nes klasifikavimo taisyklių sudarymui ir klasifikavimo kokybės vertinimui naudojama ta pati imtis [21].

Testinės imties metodą rekomenduojama taikyti, kai mokomoji imtis yra pakankamai didelė (*n≥1000*) [18]. Mokomoji imtis atsitiktinai dalinama į dvi: apmokymo ir testinę. Pirmoji naudojama klasifikavimo taisyklių sudarymui, o antroji – klaidingo klasifikavimo tikimybės įvertinimui. Rekomenduojamas testinės imties dydis - 20%-30% mokomosios imties. Jei apmokymo ir testinė imtys yra nepriklausomos, tai šiuo metodu gautas klaidingo klasifikavimo tikimybės įvertis yra nepaslinktas [28]. Metodo trūkumas – sudarant testinę imtį prarandami duomenys, kurie galėtų būti panaudoti klasifikavimo taisyklių sudarymui.

Plačiausiai paplitęs yra kryžminio patikrinimo metodas. Jis panašus į testinės imties metodą. Skiriasi tik tuo, kad naudojant kryžminio patikrinimo metodą, daug kartų sudaromos skirtingos apmokymo ir testinės imtys. Taikant atskirą šio metodo variantą, kryžminį patikrinimą išbraukiant po vieną (angl. c*ross-validation leave one out*), iš mokomosios imties išbraukiamas vienas objektas ir sudaromos klasifikavimo taisyklės, kurių pagalba klasifikuojamas išbrauktas objektas. Procesas tęsiamas kol tokiu būdu suklasifikuojami visi objektai. Tada klaidingo klasifikavimo tikimybės taškinis įvertis lygus santykiui klaidingai klasifikuotų ir visų objektų [28]. Metodo trūkumai: gauti įverčiai pasižymi didele dispersija; turint dideles imtis reikia atlikti daug skaičiavimų [4]. Plačiau visi metodo variantai (angl. *v-fold corss-validation*) pateikti literatūroje [19].

Vienas naujausių metodų yra Monte Karlo kryžminio patikrinimo. Tai tam tikra kryžminio patikrinimo metodo modifikacija. Naudojant šį metodą, mokomoji imtis atsitiktinai dalinama į testinę ir apmokymo. Tada panaudojant testinės imties metodą, įvertinama klaidingo klasifikavimo tikimybė. Procesą kartojant tam tikrą iteracijų skaičių, gaunama klaidingo klasifikavimo tikimybės taškinių įverčių imtis. Šios imties vidurkis yra klaidingo klasifikavimo tikimybės taškinis įvertis, gautas Monte Karlo kryžminio patikrinimo metodu. Metodo privalumas – gautas taškinis įvertis pasižymi mažesne dispersija nei kryžminio patikrinimo taškinis įvertis [19].

Pagrindinė įkelčių metodų idėja – sudaryti klaidingo klasifikavimo tikimybės imtį ir apskaičiuoti imties vidurkį, kuris yra ieškomas taškinis įvertis [13, 20]. Klaidingo klasifikavimo tikimybės imties sudarymui naudojama įkelčių metodika [18]. Iš n dydžio mokomosios imties ištraukiame tokio pačio dydžio imtį su pasikartojimais (tas pats objektas į imtį gali būti įtrauktas kelis kartus), kuri vadinama įkelties imtimi. Į šią imtį nepaimti objektai, priskiriami testinei imčiai. Naudojant testinės imties metodą, įvertinama klaidingo klasifikavimo tikimybė. Procesą kartojant, gaunama klaidingo klasifikavimo tikimybės įverčių imtis. Šios imties vidurkis yra klaidingo klasifikavimo tikimybės taškinis įvertis, gautas įkelčių metodu. Dažnai šis įvertis vadinamas „0.632 įkelčių“ įverčiu (angl. *„0.632 bootstrap“)* [4]. Praktikoje naudojamos kelios šio metodo modifikacijos. Vienoje iš jų prie „0.632 įkelčių“ įverčio su tam tikru svoriu pridedamas savos imties metodu gautas įvertis [19]. Įkelčių metodais gauti įverčiai pasižymi maža dispersija, tačiau gali būti paslinkti net ir didelėse imtyse [18].

1.1.10 Jautrumas, specifiškumas ir ROC kreivės

Kai turime tik dvi klases ($p=2$), pasirinkto klasifikavimo metodo kokybei įvertinti naudojami ne tik klaidingo klasifikavimo tikimybės įverčiai. Pagrindinė priežastis – bendros klaidingo klasifikavimo tikimybės įvertis neatskleidžia kaip klasifikuojami atskiros klasės objektai. Dažnai praktikoje svarbiau teisingai klasifikuoti vienos, pavyzdžiui pirmos klasės Π_1 objektus nei antros Π_2. Vienas iš pavyzdžių: ligos diagnozavimas (serga/neserga). Todėl be klaidingo klasifikavimo tikimybės įverčio naudojami klasifikavimo modelio jautrumo ir specifiškumo matai [16].

Jautrumu vadinamas teisingai klasifikuotų pirmos klasės objektų ir visų pirmos klasės objektų skaičių santykis. Specifiškumas yra teisingai klasifikuotų antros klasės objektų ir visų antros klasės objektų skaičių santykis [22]. Šiuos dydžius sieja atvirkštinė priklausomybė: didėjant jautrumui, mažėja specifiškumas, o didėjant specifiškumui, mažėja jautrumas. Tinkamesnis tas klasifikavimo modelis, kurio didesnis jautrumas, kai bendros klaidingo klasifikavimo tikimybės įverčiai vienodi.

Prie skirtingų aposteriorinės tikimybės slenkstinių reikšmių, gauname skirtingus jautrumus ir specifiškumus. Slenkstinė reikšmė yra ta, kurią viršijus objektas priskiriamas pirmai klasei, priešingu atveju – antrai. Ant x-ašies atidėjus dydį lygų 1-specifiškumas, o ant y-ašies – jautrumą braižoma ROC

kreivė (angl. *receiver operating characteristic curve*). Ji parodo ryšį tarp jautrumo ir specifiškumo. Plotas po ROC kreive naudojamas klasifikavimo metodo kokybės įvertinimui. Kuo plotas didesnis, tuo pasirinktas klasifikavimo metodas tinkamesnis. Klasifikavimo metodas laikomas puikiu, jei plotas po kreive ne mažesnis nei 0.9, geru – jei tarp 0.8 ir 0.9 [16]. Detalesnė informacija susijusi su ROC kreivėmis pateikta literatūroje [16, 22], o metodai, naudojami plotui po ROC kreive įvertinti, literatūroje [17, 27].

1.2 Klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo darbų apžvalga

Šiame skyrelyje apžvelgiami atlikti darbai, kuriuose lyginami skirtingų klasifikavimo su mokytoju metodų taikymo rezultatai. Vienintelis toks lietuvių darbas publikuotas 2001 metais yra dr. Š. Raudžio „Statistical and Neural Classifiers: An integrated approach to design“ [25]. Darbe didžiausias dėmesys skiriamas dirbtiniams neuroniniams tinklams, kurie lyginami su kitais klasifikavimo su mokytoju metodais. Analizuojami tiek paprasčiausi parametriniai metodai (Euklido atstumų klasifikatoriai, tiesinė bei kvadratinė diskrimininantinės analizės, robastinė diskriminantinė analizė), tiek ir neparametriniai (k artimiausių kaimynų, branduolinė diskriminantinės analizės, klasifikavimo medžiai ir kiti). Darbe taip pat analizuojami tinkamiausio modelio parinkimo kriterijai (įvairūs klaidingo klasifikavimo tikimybės įverčiai). MatLab programavimo kalba parašyta programa, skirta neuroninių tinklų apmokymui. Kokių programinių priemonių pagalba atliktas klasifikavimas naudojant kitus metodus nėra atskleista.

Dr. Š. Raudžio publikuotame darbe didelis dėmesys skiriamas metodų teoriniam pagrindimui. Tuo tarpu grupės autorių darbe „Machine Learning, Neural and Statistical Classification“ [18] analizuojamos metodų taikymo galimybės. Apie 20 klasifikavimo su mokytoju metodų ir įvairių jų modifikacijų naudojami klasifikuojant duomenis iš skirtingų taikymo sričių: paskolų valdymas, siekiant įvertinti klientų patikimumą ir galimybes grąžinti paskolą; atpažinimas ranka rašytų skaičių, t.y. pašto indeksų ant vokų Vokietijoje; ranka rašytų raidžių atpažinimas; galvos traumas patyrusių žmonių pasveikimo laipsnio nustatymas, širdies ligų nustatymas ir kitos sritys. Su realiais duomenimis atlikus maždaug 400 eksperimentų, gautos išvados apie metodų taikymo galimybes konkrečiose srityse. Taikant analizuojamus klasifikavimo su mokytoju metodus naudojamasi daugeliu įvairių programų, parašytų konkretiems metodams (didžioji dalis parašytos Fortran programavimo kalba). Nesukurtas patogus įrankis leidžiantis lyginti skirtingų metodų taikymo rezultatus.

Išsami metodų apžvalga pateikiama diplominiame darbe „Klassifikationsverfahren der Diskriminanzanalyse, Eine vergleichende und integrierende Übersicht“ [21]. Tiesinės, kvadratinės, diskriminantinės analizės bei klasifikavimo medžių ir neuroninių tinklų metodų taikymas atliekamas

naudojant atskiras S-Plus sistemos [15] procedūras. Darbo autoriaus S-Plus programavimo kalba parašytos procedūros skirtos savos imties, kryžminio patikrinimo bei testinės imties klaidingo klasifikavimo įverčių radimui. Tačiau visa tai nėra apjungta į vieną programą. Norint atlikti konkretų veiksmą reikia išsikviesti tam skirtą procedūrą.

T. Sueyoshi darbe "DEA-Discriminant Analysis: Methodological comparison among eight discriminant analysis approaches" [33] analizuoja naujo neparametrinės diskriminantinės analizės metodo DEA-DA (angl. *data envelopment analysis – diskriminant analysis*) taikymo galimybes. Metodo taikymo rezultatai lyginami su tiesinės, kvadratinės diskriminantinės analizės, klasifikavimo medžių bei neuroninių tinklų taikymo rezultatais.

M. Soukup, H.J. Cho ir J.K. Lee [30] klasifikavimo su mokytoju metodo parinkimui naudoja MiPP statistiką (angl. *the misclassification-penalized posterior*). Ji apskaičiuojama iš aposteriorinių tikimybių įverčių sumos atimant klaidingai klasifikuotų objektų skaičių. Nuorodos apie naudojamas programines priemones nepateiktos.

Darbe „A Comparison of Artificial Neural Networks, Logistic Regression, and Discriminant Analysis" [12] tiesinės, kvadratinės, artimiausių kaimynų bei branduolinės diskriminantinės analizės, logistinės regresinės analizės ir neuroninių tinklų metodai taikomi naudingų iškasenų paieškoje. Sprendimas apie tinkamiausią metodą priimamas atsižvelgiant į teisingai klasifikuotų objektų skaičių. Naudojamos SAS/STAT posistemės procedūros bei Nshell2 programa [36], kurios pagalba klasifikavimas atliekamas neuroninių tinklų metodu.

Tiesinės diskriminantinės analizės bei logistinės regresinės analizės taikymo rezultatai lyginami darbe „Comparison of Logistic Regression and Linear Discriminant Analysis: A Simulation Study" [23]. Priešingai nei daugelyje kitų darbų, lyginimo kriterijumi yra ne klaidingo klasifikavimo tikimybės įverčiai, o tam tikri indeksai, įvertinantys aposteriorinių tikimybių dydį. Darbe nepateikiamos nuorodos apie naudojamą programinę įrangą.

Aukščiau apžvelgtuose darbuose siekiama išsiaiškinti klasifikavimo su mokytoju metodų privalumus ir trūkumus. Tuo tarpu darbe „Prediction error estimation: a comparison of resampling methods" [19] tiriamos klasifikavimo su mokytoju metodų klasifikavimo kokybę apibūdinančių statistikų (t.y. klaidingo klasifikavimo tikimybės) vertinimo metodų savybės. Analizuojami dažniausiai naudojami taškiniai klaidingo klasifikavimo tikimybės įverčiai (testinės imties, kryžminio patikrinimo, Monte Karlo kryžminio patikrinimo, įkelčių). Aptariami įverčių privalumai ir trūkumai. Nuorodos apie naudojamas programines priemones nepateiktos.

Savos imties, „0.632 įkelčių" bei kelių variantų kryžminio patikrinimo įverčių radimo metodų savybės mažose imtyse tiriamos darbe „Is cross-validation valid for small-sample microarray classification?" [4]. Nustatyta, kad mažose imtyse klasifikavimo metodų taikymo rezultatus geriau

lyginti naudojant „0.632 įkelčių“ metodu gautus klaidingo klasifikavimo tikimybės įverčius. Darbe klaidingo klasifikavimo tikimybės vertinamos naudojant C kalba parašytas programas.

Savos imties ir kryžminio patikrinimo metodų privalumai ir trūkumai taip pat lyginami darbe „Is cross-validation better than resubstitution for ranking genes?“ [5].

Tiesinės, kvadratinės, artimiausių kaimynų diskriminantinės analizės ir klasifikavimo medžių metodų taikymo rezultatai darbe „Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data” [38] lyginami naudojant kelis kryžminio patikrinimo bei „0.632 įkelčių“ klaidingo klasifikavimo tikimybės įverčius.

Atlikus klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo literatūros apžvalgą, nustatyta, kad nėra universalaus metodo įvertinančio klasifikavimo kokybę. Todėl negalima konkrečiu atveju parinkti tinkamiausio klasifikavimo su mokytoju metodo. Apžvelgtuose darbuose [4, 5, 12, 15, 18, 19, 21, 23, 25, 30, 33, 36, 38] naudojami įvairūs klaidingo klasifikavimo tikimybės vertinimo metodai. Keliuose darbuose [4, 38] parodyta, kad daugiausia privalumų turi „0.632 įkelčių“ klaidingo klasifikavimo tikimybės vertinimo metodas, tačiau jis turi ir trūkumų. Todėl, norint konkrečiu atveju parinkti tinkamiausią metodą, skirtingų klasifikavimo su mokytoju metodų taikymo rezultatų lyginimui, rekomenduojama naudoti kelis klaidingo klasifikavimo tikimybės taškinius įverčius. Tačiau tam tikslui nei viename analizuotame darbe nesukurtas patogus programinis įrankis. Be to klasifikavimo metodų taikymo rezultatų lyginimui nenaudojami intervaliniai klaidingo klasifikavimo tikimybės įverčiai.

1.3 Klasifikavimo su mokytoju metodai sistemoje SAS

SAS sistema [28] buvo pasirinkta dėl kelių priežasčių:

- SAS sistemos licenziją turi KTU;
- SAS makrokomandų pagalba galima sukurti atskiras procedūras, skirtas konkrečiam veiksmui atlikti;
- SAS sistemoje realizuota daugiau klasifikavimo su mokytoju metodų nei kituose paketuose (SPSS [31], S-plus [15]).

Sistemoje SAS dauguma procedūrų skirtų klasifikavimui su mokytoju yra realizuotos SAS/STAT posistemėje.

Parametriniai tiesinė ir kvadratinė diskriminantinės analizės bei neparametriniai branduolinė ir artimiausių kaimynų diskriminantinės analizės yra realizuoti procedūroje *Discrim*. Taikant branduolinę diskriminantinę analizę galima pasirinkti vieną iš penkių branduolio funkcijų [28]. Realizuoti trys klaidingo klasifikavimo tikimybės vertinimo metodai (savos imties, testinės imties, kryžminio patikrinimo išbraukiant po vieną). Procedūros trūkumas – nerealizuotas automatinis glodinimo

parametro, naudojamo neparametriniuose diskriminantinės analizės metoduose, parinkimo algoritmas. Be to, vertinant tankio funkcijas, visose klasėse naudojamas tas pats glodinimo parametras, negalima nurodyti skirtingų. Nerealizuoti 1.1.4 skyrelyje aptrati robastiniai metodai.

8-oje SAS versijoje dichotominę bei politominę rangų logistinę regresinę analizę galima atlikti su *Logistc* procedūra, o politominę vardų su *Catmod* procedūra. Naujausioje 9-oje SAS versijoje visų rūšių logistinę regresinę analizę galima atlikti su *Logistic* procedūra. Pagrindiniai SAS sistemoje logistinei regresinei analizei skirtų procedūrų trūkumai yra trys:

- nevertinamos klaidingo klasifikavimo tikimybės.
- Kryžminio patikrinimo išbraukiant po vieną metodu aposteriorinių tikimybių įverčiai skaičiuojami tik kai priklausomas kintamasis yra dvireikšmis. Šiuo atveju, panaudojus gautus aposteriorinių tikimybių įverčius, galima kryžminio patikrinimo metodu įvertinti klaidingo klasifikavimo tikimybę. Kai klasių daugiau nei dvi, vartotojas šiam tikslui turi rašyti SAS programą.
- Nėra galimybės pagal sudarytas klasifikavimo taisykles automatiškai klasifikuoti testinę imtį. Tam tikslui reikia rašyti SAS programą, kuri nuskaitytų logistinių lygčių koeficientų įverčius, sudarytų klasifikavimo taisykles ir klasifikuotų testinės imties objektus.

Procedūros *Stepdisc* pagalba galima rasti tinkamiausią diskriminavimo kintamųjų rinkinį. Naudojami pažingsniniai metodai, kai kiekviename žingsnyje diskriminavimo kintamasis:

- į modelį įrašomas (angl. *forward selection*),
- iš modelio išbraukiamas (angl. *backward selection*),
- į modelį įrašomas arba išbraukiamas (angl. *stepwise selection*).

Ar diskriminavimo kintamąjį įrašyti ar išbraukti sprendžiama naudojant Vilkso statistiką [28].

Naudojant procedūrą *Candisc*, galima atlikti kanoninę diskriminantinę analizę. Čia ieškoma tiesinių diskriminavimo kintamųjų daugdarų, kurios geriausiai atskirtų klases [28]. Detaliai kanoninė diskriminantinė analizė pateikta literatūroje [7, 28].

Klasifikavimo medžių bei neuroninių tinklų metodai yra realizuoti SAS/Enterprise minner posistemėje. Klasifikavimą galima atlikti naudojant tokius klasifikavimo medžių metodus kaip CART, CHAID ir kitus [26]. Realizuoti vienasluoksniai bei daugiasluoksniai perceptronai, kurių apmokymui naudojami atbulinės sklaidos, gradientų ir kiti metodai.

Be standartinių SAS sistemoje realizuotų procedūrų galima panaudoti vartotojų sukurtas SAS makrokomandas, kurios pateikiamos SAS tinklapyje [29].

Makrokomandos *Boot* pagalba galima apskaičiuoti įvairių statistikų įverčius naudojant įkelčių metodiką. Naudojant makrokomandą *BootCI,* galima rasti klaidingo klasifikavimo tikimybės intervalinius įverčius. Intervalinių įverčių radimui naudojami keli metodai [13, 29]. Vienas iš jų, kai, įkelčių metodu sudarius klaidingo klasifikavimo tikimybės taškinių įverčių imtį, randamas klaidingo

klasifikavimo tikimybės pasikliautinasis intervalas. Abiejų makrokomandų trūkumas yra tas, kad, vertinant klaidingo klasifikavimo tikimybę, mokymo imties negalima padalinti į apmokymo ir testinę. Naudojant įkelčių metodiką, tiek taškiniai, tiek ir intervaliniai klaidingo klasifikavimo tikimybės įverčiai gaunami kai apmokymo ir testinė imtis yra ta pati.

Klasifikavimo rezultatų pateikimui grafikuose galima pasinaudoti makrokomanda *Plotit*, kuri skirta įvairių sklaidos diagramų braižymui. Ji pateikiama standartiniame SAS pakete makrokomandų bibliotekoje faile „*plotit.sas*“. Ja naudojantis galima atvaizduoti objektų suskirstymą į klases, aposteriorines tikimybes įvairiuose taškuose, galima kontroliuoti taškų spalvą, dydį, formą. Makrokomandai reikalingos SAS/STAT ir SAS/GRAPH posistemės. Išsamus šios makrokomandos aprašymas bei taikymo rekomendacijos pateikiamos literatūroje [37].

Patikrinti ar diskriminavimo kintamųjų skirstinys yra daugiamatis normalusis galima naudojant makrokomandą *Multnorm*. Suderinamumo hipotezės vienmačiu atveju patikrinimui naudojami Shapiro-Wilk W arba Kolmogorovo-Smirnovo kriterijai. Hipotezės apie daugiamatį normalųjį skirstinį tikrinimui naudojami Mardijos asimetrijos ir ekscesо koeficientų bei Henze-Zirkler kriterijai. Taip pat braižomi Mahalanobio atstumų kvadratų „Q-Q“ grafikai. Norint patikrinti suderinamumo hipotezę apie daugiamatį Gauso skirstinį reikalinga SAS/IML (neankstesnė nei 7 versija) arba SAS/ETS (neankstesnė nei 8 versija) posistemė. Jeigu tinkamos posistemės nėra, tada suderinamumo hipotezę galima patikrinti tik vienmačiu atveju. Mahalanobio atstumų kvadratų „Q-Q“ grafikų braižymui reikalinga SAS/STAT arba SAS/IML (neankstesnės nei 7 versijos). Norint gauti didelės rezoliucijos grafikus reikalinga SAS/GRAPH (neankstesnė nei 7 versija). Plačiau ši makrokomanda pateikta literatūroje [29].

ROC kreivių braižymui galima pasinaudoti makrokomanda *Rocplot*. Tačiau ji tinka tik tuo atveju, kai pasinaudojant SAS/STAT procedūra *Logistic* buvo atlikta logistinė regresinė analizė ir suformuoti ROC kreivės braižymui reikalingi duomenys. Jei klasifikavimas su mokytoju atliktas kitu metodu, tai ši makrokomanda nenaudinga.

Plotų po dviem ROC kreivėmis skirtumo įvertinimui skirta makrokomanda *Roc*. Norint pasinaudoti šia makrokomanda reikalinga SAS/Base ir SAS/IML posistemės (ne ankstesnės nei 7 versijos). Plačiau makrokomandos *Roc* ir *Rocplot* pateiktos literatūroje [29].

Atlikus SAS sistemos klasifikavimo su mokytoju metodus realizuojančių procedūrų ir makrokomandų apžvalgą, nustatyta, kad nėra programinio įrankio leidžiančio patogiai lyginti klasifikavimo su mokytoju metodų taikymo rezultatus. Tam reikia naudoti skirtingas procedūras bei makrokomandas, kurios pateikia rezultatus įvairiais formatais bei skirtingose duomenų žingsnio programos vietose. Vartotojas turi gerai žinoti SAS programavimo kalbą bei turėti pakankamus programavimo įgūdžius. Todėl klasifikavimo rezultatų lyginimas reikalauja didelių darbo bei laiko sąnaudų ir aukštos kvalifikacijos programavimo srityje. Dar vienas trūkumas yra tas, kad keliose SAS

procedūrose realizuoti tik du arba visai nerealizuoti (logistinės regresinės analizės atveju) klaidingo klasifikavimo tikimybės vertinimo metodai, taip pat visose procedūrose nerealizuoti klaidingo klasifikavimo tikimybės intervalinių įverčių radimo metodai.

1.4 Darbe sprendžiami uždaviniai

Pagrindinis darbo tikslas – išplėsti SAS sistemos galimybes klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankiu ir atlikti pasirinktų duomenų klasifikavimą.

Sprendžiami uždaviniai:

- susipažinti su parametriniais ir naujausiais neparametriniais klasifikavimo su mokytoju metodais bei jų taikymo rezultatų lyginimo kriterijais.
- Parengti įvairių klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo metodiką ir pasiūlyti programinių priemonių, automatizuojančių šio uždavinio sprendimą, realizavimo principus.
- Išplėsti SAS sistemos galimybes sukuriant makrokomandų rinkinį, kuris realizuotų tiesinės, kvadratinės, branduolinės, artimiausių kaimynų diskriminantinės analizės bei logistinės regresijos metodų klasifikavimo kokybės lyginimą naudojant klaidingo klasifikavimo tikimybės taškinius įverčius, gautus savos imties, kryžminio patikrinimo, įkelčių bei Monte Karlo kryžminio patikrinimo metodais, taip pat intervalinius įverčius, gautus įkelčių metodika.
- Parengti sukurto programinio įrankio taikymo rekomendacijas bei apribojimus.
- Panaudojus sukurtą įrankį atlikti pasirinktų metodų klasifikavimo galimybių lyginimą pasirinktiems duomenims.

2 TIRIAMOJI DALIS

2.1 Klasifikavimo metodų taikymo rezultatų lyginimo metodika

Šiame skyrelyje aprašoma klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo metodika. Poskyryje 2.1.1 nurodyti pasirinkti klasifikavimo su mokytoju metodai bei jų pasirinkimo kriterijai, poskyryje 2.1.2 – klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo kriterijai, o poskyryje 2.1.3 – pasirinkti parametrinių klasifikavimo metodų taikymo prielaidų tikrinimo kriterijai.

2.1.1 Klasifikavimo metodai

Kadangi pagrindinis darbo tikslas yra išplėsti žinomos statistinės duomenų analizės sistemos SAS galimybes klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankiu, pasirinkti SAS/STAT posistemėje realizuoti klasifikavimo su mokytoju metodai:

- tiesinė diskriminantinė analizė (TDA),
- kvadratinė diskriminantinė analizė (KDA),
- branduolinė diskriminantinė analizė (BDA),
- artimiausių kaimynų diskriminantinė analizė(AKDA),
- politominė vardų logistinė regresija (LR).

Kiti metodai, tokie kaip klasifikavimo medžiai bei neuroniniai tinklai, yra realizuoti SAS/Enterprise Minner modulyje, kurio licenzijos neturi KTU. Todėl šiame darbe jie neanalizuojami.

Atliekant branduolinę diskriminantinę analizę naudojamos branduolio funkcijos:

- tolygioji $$K(u)=\frac{1}{2}I(|u|\le 1), \quad \textbf{(2.1)}$$
- normalioji $$K(u)=\frac{1}{2\pi}\exp\left(-\frac{1}{2}u^2\right), \quad \textbf{(2.2)}$$
- Epanechnikovo $$K(u)=\frac{3}{4}(1-u^2)I(|u|\le 1), \quad \textbf{(2.3)}$$
- dvisvorė $$K(u)=\frac{15}{16}(1-u^2)^2 I(|u|\le 1), \quad \textbf{(2.4)}$$
- trisvorė $$K(u)=\frac{35}{32}(1-u^2)^3 I(|u|\le 1), \quad \textbf{(2.5)}$$

čia $I(|u|\le 1)=\begin{cases}1, |u|\le 1\\ 0, |u|>1\end{cases}$. Branduolio funkcijos pateiktos vienmačiu atveju, t.y. kai *u* skaliarinis dydis. Daugiamatės funkcijos pateiktos literatūroje [28].

Taikant branduolinę ir artimiausių kaimynų diskriminantines analizes, reikia parinkti glodinimo parametrą, kuris nurodomas klasifikavimo procedūroje. Literatūroje [11 ,21, 28] pasiūlyti keli optimalaus glodinimo parametro parinkimo metodai, bet nei vienas iš jų nerealizuotas SAS procedūrose, skirtose klasifikavimui su mokytoju atlikti. Šio uždavinio sprendimui nėra parašytų ir SAS makrokomandų. Optimalaus glodinimo parametro, kaip ir branduolio funkcijos, parinkimo uždavinys šiame darbe nesprendžiamas.

2.1.2 Klasifikavimo metodų taikymo rezultatų lyginimo kriterijai

Pagrindinis metodų taikymo rezultatų lyginimo kriterijus – klaidingo klasifikavimo tikimybės įverčiai. Atlikus literatūros analizę 1.3 skyriuje klaidingo klasifikavimo tikimybės vertinimui atrinkti keturi metodai:

- savos imties (SI) (angl. *resubstition*),
- kryžminio patikrinimo išbraukiant po vieną (KP) (angl. *cross-validation leave one out*),
- įkelčių (IK) (angl. *bootstrap*),
- Monte Karlo kryžminio patikrinimo (MKKP) (angl. *Monte Carl cross-validation*).

Visų keturių metodų veikimo principai yra panašūs (2 pav.). Pirmiausia mokomoji imtis dalinami į apmokymo ir testinę imtį. Pirmoji iš jų naudojama klasifikavimo taisyklių sudarymui. Tada panaudojant sudarytas taisykles klasifikuojami testinės imties objektai. Skaičiuojama kiek kiekvienoje klasėje yra klaidingai klasifikuotų objektų. Klaidingo klasifikavimo tikimybių įvertinimui naudojamos formulės:

$$\hat{\alpha}_i = \frac{e_i}{n_i}, \tag{2.6}$$

$$\hat{\alpha} = \frac{\sum_i e_i}{n}, \tag{2.7}$$

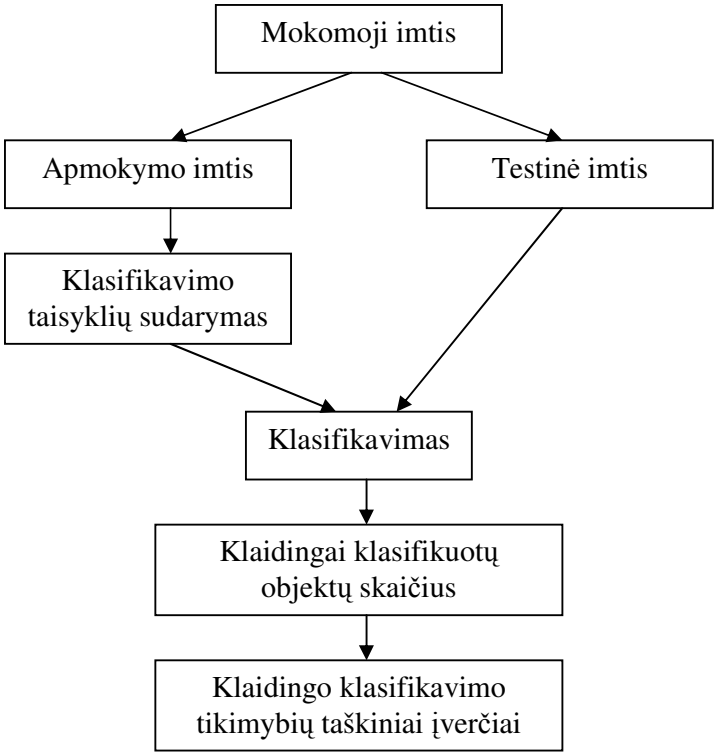
čia $\hat{\alpha}_i$ – i-tosios klasės klaidingo klasifikavimo tikimybės įvertis, $\hat{\alpha}$ – bendros klaidingo klasifikavimo tikimybės įvertis, e_i – i-tosios klasės klaidingai klasifikuotų objektų skaičius, n_i – i-tosios klasės objektų skaičius, n – testinės imties dydis.

Taikant skirtingus klaidingo klasifikavimo tikimybės vertinimo metodus, skiriasi tik pirmasis klaidingo klasifikavimo tikimybės vertinimo etapas, t.y. mokomosios imties dalinimas į apmokymo ir testinę imtis.

Taikant SI metodą, apmokymo ir testinė imtis yra ta pati.

Taikant KP metodą, iš mokomosios imties išbraukiamas vienas objektas, kuris priskiriamas testinei imčiai, o visi likusieji – apmokymo imčiai. Atlikus klasifikavimą, iš mokomosios imties

išbraukiamas kitas objektas, kuris vėl priskiriamas testinei imčiai, o visi likusieji – apmokymo. Procesas kartojamas, kol visi objektai suklasifikuojami. Tuomet pagal 2.6 ir 2.7 formules vertinamos klaidingo klasifikavimo tikimybės. Pasirinktas šis kryžminio patikrinimo metodo atvejis, nes jis realizuotas SAS/STAT procedūroje *Discrim*.



2 pav. Klaidingo klasifikavimo tikimybės vertinimo metodika

Taikant IK metodą, sudaroma įkelties imtis, t.y. iš *n* dydžio mokomosios imties objektų sudaroma tokio pačio dydžio imtis su pasikartojimais (tas pats objektas į imtį gali būti įtrauktas kelis kartus). Įkelties imtis naudojama kaip apmokymo, o į šią imtį neįtraukti objektai priskiriami testinei imčiai. Testinės imties metodu vertinama klaidingo klasifikavimo tikimybė (naudojama 2.7 formulė). Kartojant procesą sudaroma klaidingo klasifikavimo tikimybės įverčių imtis. Gautosios imties vidurkis yra taškinis klaidingo klasifikavimo tikimybės įvertis, gautas IK metodu. Analogiškai randami ir klasių klaidingo klasifikavimo tikimybių įverčiai.

Taikant MKKP metodą, iš mokomosios imties atsitiktinai be pasikartojimų išrenkame nurodytą skaičių *t* objektų ir juos priskiriame testinei imčiai, o likusius – apmokymo. Atlikus klasifikavimą, pagal 2.7 formulę įvertinama klaidingo klasifikavimo tikimybė. Procesą kartojant nurodytą iteracijų skaičių, gaunama klaidingo klasifikavimo tikimybės įverčių imtis. Suradę imties vidurkį, gauname

taškinį Monte Karlo kryžminio patikrinimo klaidingo klasifikavimo tikimybės įvertį. Analogiškai randami ir klasių klaidingo klasifikavimo tikimybių įverčiai.

Pagrindinis MKKP ir IK metodų, skirtumas yra tas, kad taikant MKKP metodą apmokymo imtys sudaromos be pasikartojimų, o IK metodą – su pasikartojimais. Antruoju atveju galima rasti ne tik taškinį klaidingo klasifikavimo tikimybės įvertį, bet ir intervalinį. Literatūroje [13] pateikiami keli tokio įverčio radimo metodai. Visi šie metodai bendrai vadinami įkelčių metodais, nes bet kuriuo metodu ieškant klaidingo klasifikavimo tikimybės intervalinio įverčio, naudojama klaidingo klasifikavimo tikimybės taškinių įverčių imtis, gauta įkelčių metodu. Šiame darbe pasirinktas paprasčiausias neparametrinis procentilių metodas, pagal kurį klaidingo klasifikavimo tikimybės 95% pasikliautinojo intervalo rėžiai atitinkamai lygūs 2.5-tam ir 97.5-tam klaidingo klasifikavimo tikimybės įverčių imties procentiliui.

Klasifikavimo su mokytoju metodų taikymo rezultatų lyginimui taip pat naudosime klaidingo klasifikavimo tikimybės įverčių imties, gautos IK metodu, bei klaidingo klasifikavimo tikimybės įverčių imties, gautos MKKP metodu, medianas ir jų 95% pasikliautinuosius intervalus. Imties pasikliautinieji intervalai vertinami neparametriniu Hahn ir Meeker metodu, realizuotu SAS sistemos procedūroje *Univariate* [28].

Klasifikavimo su mokytoju metodų taikymo rezultatų lyginimui naudosime ne tik klaidingo klasifikavimo tikimybės įverčių skaitines reikšmes, bet ir grafikus:

- klaidingo klasifikavimo tikimybės įverčių imties, gautos IK metodu, ir klaidingo klasifikavimo tikimybės įverčių imties, gautos MKKP metodu, stačiakampes diagramas su išpjovomis.
- Klasifikavimo rezultatų grafikus, kuriuose pavaizduotos sudarytos objektų klasės (braižomi tik dvimačiu atveju (kai $p=2$)).

Dviejų klasių atveju jautrumas ir specifiškumas atskirai neskaičiuojami, nes jie atitinkamai lygūs $1-\hat{\alpha}_1$ ir $1-\hat{\alpha}_2$ ($\hat{\alpha}_1, \hat{\alpha}_2$ – SI metodu gauti įverčiai).

2.1.3 Klasifikavimo metodų prielaidų tikrinimas

Tiek tiesinę, tiek ir kvadratinę diskriminantines analizes galima taikyti, kai kiekvienoje klasėje p-mačio diskriminavimo kintamojo skirstinys yra daugiamatis normalusis. Ši sąlyga tikrinama pagal literatūroje [29] pateiktą metodiką.

- Tikrinamos suderinamumo hipotezės kiekvienam vienmačiam diskriminavimo kintamajam atskirai.
- Tikrinama hipotezė apie daugiamačio skirstinio asimetrijos koeficientą. Naudojamas Mardijos asimetrijos koeficiento kriterijus.

- Tikrinama hipotezė apie daugiamačio skirstinio eksceso koeficientą. Naudojamas Mardijos eksceso koeficiento kriterijus.
- Tikrinama suderinamumo hipotezė daugiamačiu atveju naudojant Henze-Zirkler kriterijų.
- Braižomi Mahalanobio atstumų kvadratų teorinio ir empirinio skirstinių kvantilių lyginimo „Q-Q“ grafikai.

Pirmiausia su reikšmingumo lygmeniu $\alpha = 0.05$ tikrinama nulinė hipotezė, kad vienmatis diskriminavimo kintamojo skirstinys yra normalusis su vidurkiu ir standartiniu nuokrypiu, kurie atitinkamai lygūs imties vidurkiui ir standartiniam nuokrypiui. Kai $n_i < 2000$ (n_i i-tosios klasės objektų skaičius), naudojamas Shapiro-Wilk W kriterijus, priešingu atveju – Kolmogorovo-Smirnovo. Taikant pirmąjį, naudojama Shapiro-Wilk W statistika [28]:

$$W = \frac{\left(\sum_{k=1}^{n_i} a_k X_{(k)}\right)^2}{\sum_{k=1}^{n_i} (X_k - \hat{\mu}_i)^2}, \tag{2.8}$$

čia $X_{(k)}$ – k-tasis variacinės eilutės $X_{(1)} < X_{(2)} < ... < X_{(n_i)}$ elementas, $a = (a_1, a_2, ..., a_{n_i})$ – konstantos, sugeneruotos pagal n_i-matį normalųjį skirstinį [28]. Atlikus statistikos W transformaciją [28], gauname statistiką, kurios skirstinys yra standartinis normalusis.

Taikant Kolmogorovo-Smirnovo kriterijų, vertinamas empirinės ir teorinės pasiskirstymo funkcijos didžiausias nuokrypis, t.y. statistika:

$$D = \sup_{x \in \Pi_i} | \hat{F}(x) - F(x) |, \tag{2.9}$$

čia $\hat{F}(x)$ – i-tosios klasės empirinė, o $F(x)$ – teorinė pasiskirstymo funkcijos. Statistikos D reikšmė d gaunama imant imties konkrečios klasės objektų variacinę eilutę ir apskaičiuojant $d^+ = \max_{1 \le k \le n_i} \left(\frac{k}{n_i} - F_0(x_{(k)}) \right)$, $d^- = \max_{1 \le k \le n_i} \left(F_0(x_{(k)}) - \frac{k-1}{n_i} \right)$, $d = \max(d^+, d^-)$. Gautoji statistikos reikšmė lyginama su Kolmogorovo skirstinio kvantiliu [28].

Taikant Mardijos asimetrijos koeficiento kriterijų, naudojama statistika:

$$B_1 = \frac{1}{n^2} \sum_{i=1}^{n} \left((X_i - \hat{\mu})^T \Sigma^{-1} (X_i - \hat{\mu}) \right)^3. \tag{2.10}$$

Mardija įrodė, kad $\frac{n}{6} B_1 \sim \chi^2 \left(p(p+1)(p+2)/6 \right)$ [3, 28].

Taikant Mardijos eksceso koeficiento kriterijų, naudojama statistika:

$$B_2 = \frac{1}{n} \sum_{i=1}^{n} \left((X_i - \hat{\mu})^T \Sigma^{-1} (X_i - \hat{\mu}) \right)^2. \tag{2.11}$$

Mardija įrodė, kad $B_2 \sim N\left(p(p+2); \sqrt{8p(p+2)/n}\right)$ [3, 28].

Taikant Henze-Zirkler kriterijų, naudojama statistika $T_\beta(p)$ [3, 28]:

$$T_\beta(p) = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n}\exp\left(-\frac{\beta^2}{2}\left(X_j - X_k\right)^T \hat{\Sigma}^{-1}\left(X_j - X_k\right)\right) - \tag{2.12}$$

$$-\frac{2}{n}\left(1+\beta^2\right)^{-p/2}\sum_{j=1}^{n}\exp\left(-\frac{\beta^2}{2\left(1+\beta^2\right)}\left(X_j - \hat{\mu}\right)^T \hat{\Sigma}^{-1}\left(X_j - \hat{\mu}\right)\right) + \left(1+2\beta^2\right)^{-p/2},$$

čia $\beta = \frac{1}{\sqrt{2}}\left(\frac{2p+1}{4}\right)^{1/(p+4)} \cdot n^{1/(p+4)}$. $T_\beta(p)$ pasiskirsčius pagal lognormalųjį skirstinį su vidurkiu ir dispersija, kurie atitinkamai apskaičiuojami: $\mu = 1 - \left(1+2\beta^2\right)^{-p/2} \cdot \left(1 + \frac{p\beta^2}{1+2\beta^2} + \frac{p(p+2)\beta^4}{2(1+2\beta^2)^2}\right)$,

$\sigma^2 = 2\left(1+4\beta^2\right)^{-p/2} + 2\left(1+2\beta^2\right)^{-p} \cdot \left(1 + \frac{2p\beta^4}{(1+2\beta^2)^2} + \frac{3p(p+2)\beta^8}{4(1+2\beta^2)^4}\right) - 4w^{-p/2}\left(1 + \frac{3p\beta^4}{2w} + \frac{p(p+2)\beta^8}{2w^2}\right)$

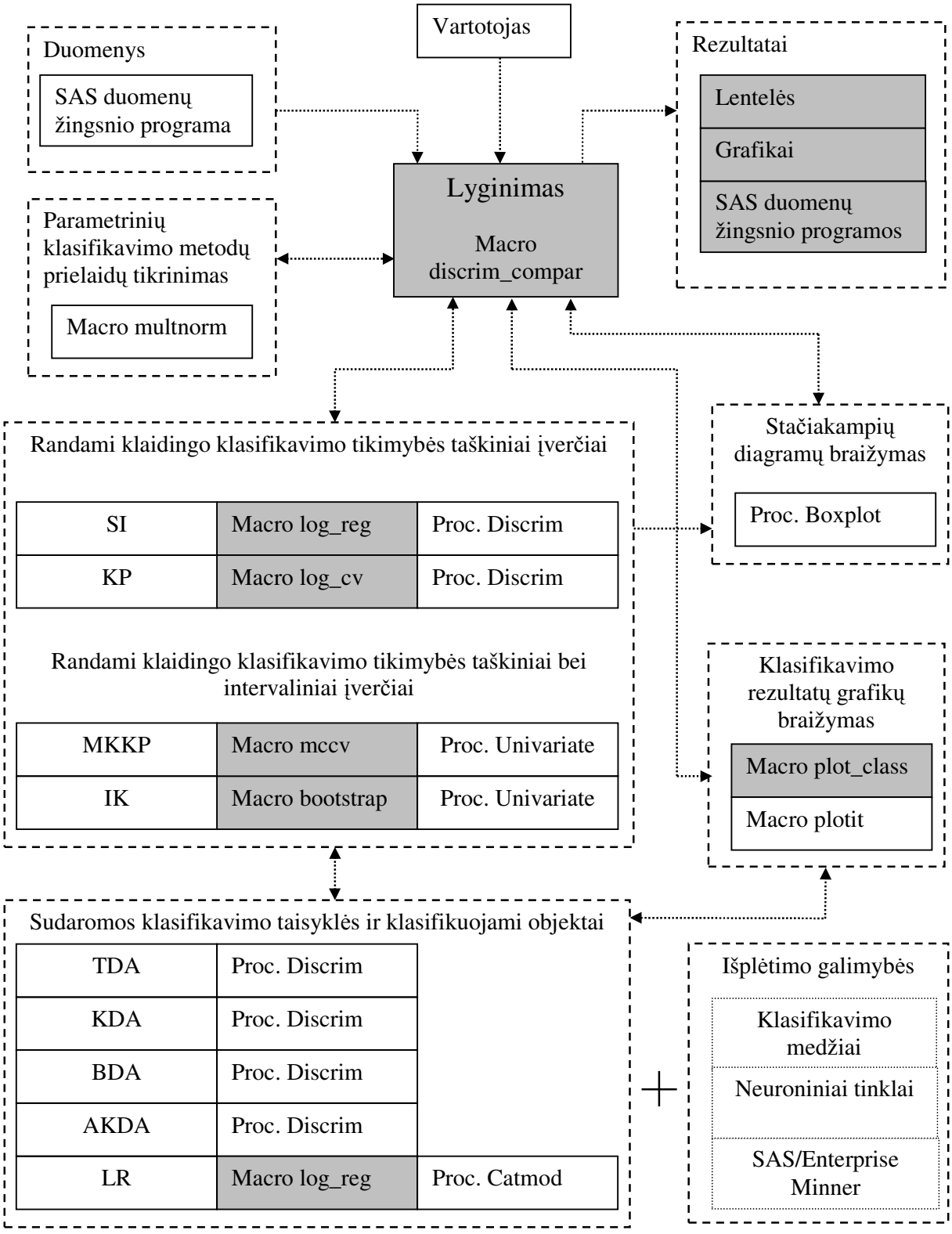
$w = (1+\beta^2)(1+3\beta^2)$ [34].

Naudojant Mahalanobio atstumų kvadratų „Q-Q“ grafikus, galima tik vizualiai spręsti apie nukrypimus nuo daugiamačio normaliojo skirstinio. „Q-Q“ grafike lyginami objektų atstumų iki klasės centro empirinės skirstinio funkcijos kvantiliai su teorinio skirstinio kvantiliais. Naudojamas kvadratinis Mahalanobio atstumo matas. Yra nustatyta, kad didelėse imtyse objektų atstumai iki klasės centro yra pasiskirstę pagal Chi kvadrato skirstinį su *p* laisvės laipsnių. Todėl šiuo atveju teoriniu skirstiniu yra $\chi^2(p)$. Plačiau šis grubus kriterijus pateikiamas literatūroje [3].

Taikant tiesinę diskriminantinę analizę, reikia patikrinti ar skirtingose klasėse kovariacijų matricos yra vienodos. Homogeniškumo hipotezės tikrinimui naudojamas Barleto kriterijus, realizuotas SAS sistemos procedūroje *Univariate* [28].

2.2 Metodų taikymo rezultatų lyginimo įrankis

Programinis klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankis – tai SAS makrokomandų rinkinys, kurį sudaro 30 autoriaus parašytų makrokomandų ir dvi SAS sistemos makrokomandos (*plotit* ir *multnorm*) Jos visos apjungtos vienoje pagrindinėje makrokomandoje *discrim_compar*. Eksperimentų atlikimui naudojamos trys papildomos, autoriaus parašytos makrokomandos, skirtos pradinių duomenų generavimui ir glodinimo parametro parinkimui [1 priedas]. Visų darbe parašytų makrokomandų apimtis – 2825 eilutės SAS programavimo kalba. Jų tekstai pateikti prie darbo pridėto kompaktinio disko faile „discrim_compar.sas“, o SAS sistemos makrokomandų *plotit* ir *multnorm* tekstai pateikti atitinkamai failuose „plotit.sas“ ir „multnorm.sas“.



3 pav. Sukurto programinio įrankio struktūra

Sukurto programinio įrankio (pagrindinės makro komandos *discrim_compar*) struktūra pavaizduota paveiksle 3 pav., kuriame prie kiekvieno veiksmo ar metodo pateikta kokios procedūros ar makrokomandos tą veiksmą ar metodą realizuoja. Tamsiau pavaizduotuose langeliuose pateiktos

autoriaus parašytos makrokomandos, o baltuose langeliuose – SAS sistemos procedūros arba kitų SAS vartotojų sukurtos makrokomandos. Kai kurie veiksmai atliekami naudojant kelias procedūras ar makrokomandas. Pvz., sudarant klasifikavimo taisykles ir klasifikuojant objektus LR metodu, naudojama SAS/STAT procedūra *catmod* ir autoriaus parašyta makrokomanda *log_reg*.

Vartotojas, atliekantis klasifikavimo su mokytoju metodų taikymo rezultatų lyginimą, dirba tik su makrokomanda *discrim_compar*. Visos kitos lyginimui reikalingos procedūros ir makrokomandos iškviečiamos automatiškai. Vartotojui reikia nurodyti tik SAS duomenų žingsnio programą, kurioje saugomi pradiniai duomenys, t.y. mokomoji imtis.

Vertinant klaidingo klasifikavimo tikimybę SI metodu, mokomoji imtis siunčiama klasifikavimo taisyklių sudarymo procedūroms bei makrokomandoms. Naudojant mokomosios imties objektus ir pasirinktus metodus, sudaromos klasifikavimo taisyklės, kurias naudojant klasifikuojami tos pačios imties objektai. Suklasifikuota imtis grąžinama atgal makrokomandoms bei procedūroms, vertinančioms klaidingo klasifikavimo tikimybę. Tuomet skaičiuojami kiekvienos klasės klaidingai klasifikuoti objektai bei vertinamos kiekvienos klasės bei bendra klaidingo klasifikavimo tikimybės.

Naudojant KP metodą, mokomoji imtis dalinama į apmokymo ir testinę, kurios siunčiamos klasifikavimo taisyklių sudarymo ir klasifikavimo procedūroms bei makrokomandoms. Naudojant apmokymo imtį ir pasirinktus metodus, sudaromos klasifikavimo taisyklės bei klasifikuojami testinės imties objektai. Suklasifikuota testinė imtis grąžinama atgal klaidingo klasifikavimo tikimybę vertinančioms procedūroms ir makrokomandoms. Mokomoji imtis vėl dalinama į apmokymo ir testinę sudaromos klasifikavimo taisyklės ir klasifikuojami objektai. Procesas kartojamas, kol suklasifikuojami visi mokomosios imties objektai. Tuomet SI metodo procedūros ir makrokomandos įvertina kiekvienos klasės bei bendrą klaidingo klasifikavimo tikimybes.

Naudojant MKKP ir IK metodus, priklausomai nuo metodo mokomoji imtis dalinama į apmokymo ir testinę. Šios imtys siunčiamos klasifikavimo taisyklių sudarymo ir objektų klasifikavimo procedūroms bei makrokomandoms. Naudojant apmokymo imtį ir pasirinktus metodus, sudaromos klasifikavimo taisyklės bei klasifikuojami testinės imties objektai. Suklasifikuota testinė imtis grąžinama atgal klaidingo klasifikavimo tikimybės vertinimo procedūroms ir makrokomandoms, kurios vertina kiekvienos klasės bei bendrą klaidingo klasifikavimo tikimybes. Procesą kartojant nurodytą iteracijų skaičių, sugeneruojamos (atitinkamai MKKP ir IK metodais) klaidingo klasifikavimo tikimybės įverčių imtys, kurias naudojant klaidingo klasifikavimo tikimybės vertinimo procedūros ir makrokomandos skaičiuoja:

- imčių vidurkius, kurie yra atitinkamų metodų (MKKP ir IK) kiekvienos klasės bei bendros klaidingo klasifikavimo tikimybių įverčiai;
- imties, gautos IK metodu, 2.5-tą ir 97.5-tą procentilius, kurie yra atitinkami klaidingo klasifikavimo tikimybės 95% pasikliautinojo intervalo rėžiai;

- imčių medianas bei medianų 95% pasikliautinojo intervalus.

Gauti klaidingo klasifikavimo tikimybių įverčiai siunčiami pagrindinei makrokomandai, kurioje jie surašomi į atitinkamas lenteles ir pateikiami vartotojui.

Jeigu makrokomandai *discrim_compar* vartotojas nurodo braižyti klaidingo klasifikavimo tikimybės įverčių imties stačiakampes diagramas su išpjovomis, tai klaidingo klasifikavimo tikimybę vertinančių procedūrų ir makrokomandų MKKP ir IK metodais sugeneruotos klaidingo klasifikavimo tikimybės įverčių imtys siunčiamos procedūrai *boxplot*, kuri braižo stačiakampes diagramas. Gauti grafikai pateikiami vartotojui.

Jeigu makrokomandai *discrim_compar* vartotojas nurodo braižyti klasifikavimo rezultatų grafikus, tai sudaroma objektų matrica, kuri kaip testinė imtis siunčiama klasifikavimo taisyklių sudarymo bei objektų klasifikavimo makrokomandoms ir procedūroms. Naudojant mokomąją imtį ir pasirinktus klasifikavimo su mokytoju metodus, sudaromos klasifikavimo taisyklės ir klasifikuojami testinės imties objektai. Suklasifikuota testinė imtis siunčiama rezultatų grafikų braižymo makrokomandai *plot_class*. Ši makrokomanda paruošia duomenis (uždeda reikiamus formatus) ir iškviečia makrokomandą *plotit*, kuri nubraižo grafikus. Gauti grafikai pateikiami vartotojui.

Jeigu vartotojas nurodo makrokomandai *discrim_compar* tikrinti parametrinių klasifikavimo su mokytoju metodų prielaidas, iškviečiama makrokomanda *multnorm*. Vartotojui pateikiami suderinamumo bei homogeniškumo hipotezių tikrinimo rezultatai ir Mahalanobio atstumų kvadratų „Q-Q“ grafikai.

Pagrindinę makrokomandą *discrim_compar* galima papildyti naujais klasifikavimo su mokytoju metodais, pvz., klasifikavimo medžiais ar neuroniniais tinklais, realizuotais SAS/Enterprise Minner modulyje. Reikia tik paruošti procedūras, kurios naudodamos nurodytus pradinius duomenis (apmokymo imtį), sudarytų klasifikavimo taisykles ir išvestų rezultatus (suklasifikuotus testinės imties objektus). Įrankis gali būti išplėstas ir naujais klasifikavimo su mokytoju metodų taikymo lyginimo kriterijais.

2.3 Vartotojo sąsaja

Sukurtas klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankis skirtas darbui SAS sistemoje programavimo režime. Vartotojas norėdamas atlikti lyginimą turi iškviesti makrokomandą *discrim_compar* ir nurodyti šios makrokomandos parametrus.

Pateikiama makrokomandos antraštė:

```
%discrim_compar   (     d_in=,                  resub=1,
                        variables=,             cv=0,
                        class=,                 bootstrap_iter=0,
                        class_num=0,            mccv_iter=0,
                                                t=0,


                        LDA=1,                  pp=0,
                        QDA=1,                  graphs=0,
                        KDA_biw=1,              box_plot=0,
                        r_biw=1,                intermediate_rez=0,
                        KDA_epa=1,              assumption_check=1,
                        r_epa=1,                english=0,
                        KDA_norm=1,             print=1,
                        r_norm=1,               pp_out=pp_out,
                        KDA_tri=1,              pp_out_test=pp_out_test,
                        r_tri=1,                resub_out=resub_errors,
                        KDA_uni=1,              cv_out=cv_errors,
                        r_uni=1,                bootstrap_out=bootstrap_errors,
                        NNDA=1,                 mccv_out=mccv_errors,
                        nn=1,                   errors=errors,
                        LR=1,                   median=median,
                                                class_name=class_       ).
```

Visi *discrim_compar* parametrai suskirstyti į keturias grupes:

- pradinių duomenų ir kintamųjų parametrai,
- klasifikavimo su mokytoju metodų parametrai,
- klaidingo klasifikavimo tikimybės vertinimo metodų parametrai,
- rezultatų išvedimo parametrai.

Vartotojas privalo nurodyti tik pradinių duomenų ir kintamųjų parametrus. Likę parametrai nurodomi tik tuo atveju, kai norima pakeisti pagal nutylėjimą priskirtą reikšmę. Nurodant parametrą reikia įvesti parametro vardą, parašyti lygybės ženklą ir norimą parametro reikšmę. Po parametro reikšmės dedamas kablelis, išskyrus paskutinio nurodyto parametro reikšmę. Parametrai nurodantys, ar atlikti konkretų veiksmą ar ne, turi dvi reikšmes: 0 reiškia, kad veiksmas nebus atliekamas, 1 – veiksmas bus atliekamas. Pvz., *pint=0* reiškia, kad rezultatai nebus spausdinami, o *pint=1* – rezultatai bus spausdinami.

Pradinių duomenų ir kintamųjų parametrai skirti makrokomandai nurodyti mokomąją imtį, priklausomą bei diskriminavimo kintamuosius:

- *d_in* – SAS duomenų žingsnio programa, kurioje saugomi priklausomo ir diskriminavimo kintamųjų reikšmės.
- *variables* – diskriminavimo kintamųjų vardai. Tarp vardų dedamas tarpas. Diskriminavimo kintamieji turi būti išmatuoti intervalų skalėje.

- *class* – priklausomojo kintamojo vardas. Klasės turi būti užkoduotos natūriniais skaičiais: pirmoji klasė = 1, antroji = 2 ir t.t.
- *class_num = 0* – klasių skaičius. Jei parametro reikšmė yra 0, tai makrokomanda skaičiuoja kiek yra klasių. Vartotojas gali pats nurodyti klasių skaičių, tokiu atveju bus sutaupytas laikas, nes nereiks skaičiuoti klasių.

Klasifikavimo metodų parametrais nurodoma ar atlikti klasifikavimą naudojant konkretų klasifikavimo su mokytoju metodą. Nurodžius klasifikavimo su mokytoju metodą, nurodomas to metodo parametras (tik branduolinės ir artimiausių kaimynų diskriminantinių analizių atveju). Naudojami klasifikavimo metodų parametrai:

- *LDA* – tiesinė diskriminantinė analizė,
- *QDA* – kvadratinė diskriminantinė analizė,
- *KDA_biw* – branduolinė diskriminantinė analizė su dvisvoriu branduoliu,
- *R_biw* – BDA metodo su dvisvoriu branduoliu glodinimo parametras,
- *KDA_epa* – branduolinė diskriminantinė analizė su Epanechnikovo branduoliu,
- *R_epa* – BDA metodo su Epanechnikovo branduoliu glodinimo parametras,
- *KDA_norm* – branduolinė diskriminantinė analizė su normaliuoju branduoliu,
- *R_norm* – BDA metodo su normaliuoju branduoliu glodinimo parametras,
- *KDA_tri* – branduolinė diskriminantinė analizė su trisvoriu branduoliu,
- *R_tri* – BDA metodo su trisvoriu branduoliu glodinimo parametras,
- *KDA_uni* – branduolinė diskriminantinė analizė su tolygiuoju branduoliu,
- *R_uni* – BDA metodo su tolygiuoju branduoliu glodinimo parametras,
- *NNDA* – artimiausių kaimynų diskriminantinė analizė,
- *Nn* – AKDA metodo glodinimo parametras,
- *LR* – politominė vardų logistinė regresinė analizė.

Klaidingo klasifikavimo tikimybės vertinimo metodų parametrais nurodomi klaidingo klasifikavimo tikimybės vertinimui naudojami metodai. Nurodant MCCV, IK, metodus reikia nurodyti iteracijų skaičių. Jei iteracijų skaičius 0, tai metodas nenaudojamas.

- *Resub* – savos imties metodas,
- *Cv* – kryžminio patikrinimo išbraukiant po vieną metodas,
- *Bootstrap_iter* – įkelčių metodo iteracijų skaičius,
- *Mccv_iter* – Monte Karlo kryžminio patikrinimo metodo iteracijų skaičius,
- *t* – nurodo kiek procentų mokomosios imties priskirti testinei imčiai, kai kiekvienoje iteracijoje mokomoji imtis dalinama į apmokymo ir testinę. Galimos reikšmės {0, 1,..., 100}.

Jei *t=0*, tai kiekvienoje iteracijoje testinei imčiai priskiriamas atsitiktinis skaičius objektų $t \sim T(1; n/2)$.

Visi metodų taikymo lyginamosios analizės rezultatai išvedami vartotojui patogiu būdu. Pirmiausia yra išvedamos mokomosios imties skaitinės charakteristikos (vidurkis, dispersija ir kt.) Vartotojas turi galimybę nurodyti kokias sukurti SAS duomenų žingsnio programas, į kurias surašomi metodų lyginimui naudojamų kriterijų reikšmės. Skirtingus metodus apibūdinančių kriterijų reikšmės išvedamos vienoje lentelėje, kad vartotojui būtų patogu juos lyginti. Kai yra du diskriminavimo kintamieji, pateikiami klasifikavimo rezultatų grafikai, kuriuose pavaizduotos objektų klasės. Rezultatų išvedimas valdomas naudojant parametrus:

- *Pp* – mokomosios imties objektų aposteriorinių tikimybių išvedimas.
- *Graphs* – nurodo ar braižyti taškų sklaidos grafikus.
- *Box_plot* – nurodo ar braižyti klaidingo klasifikavimo tikimybės įverčių imties stačiakampes diagramas su išpjovomis.
- *Intermediate_rez* – nurodo ar išvesti tarpinius metodų rezultatus, t.y. LR lygčių koeficientų įverčius, klasifikavimo rezultatų lenteles, klaidingai klasifikuotų objektų lenteles, kiekvienos klasės klaidingo klasifikavimo tikimybės įverčių, gautų SI, KP, MKKP, IK metodais, lenteles.
- *Assumption_check* – nurodo ar tikrinti parametrinių metodų prielaidas.
- *English* – nurodo anglų (english = 1) ar lietuvių (english = 0) kalba išvesti pagrindinius rezultatus. Tarpiniai rezultatai išvedami anglų kalba.
- *Print* – nurodo ar spausdinti rezultatus.
- *Pp_out* – vardas duomenų žingsnio programos, kurioje saugomos mokomosios imties objektų aposteriorinės tikimybės.
- *Pp_out_test* – vardas duomenų žingsnio programos, kurioje saugomos grafikų braižymui naudojamų objektų aposteriorinės tikimybės. Duomenų žingsnio programa sukuriama, kai *graph=1* ir yra du diskriminavimo kintamieji.
- *Resub_out* – vardas duomenų žingsnio programos, kurioje saugomos kiekvienos klasės klaidingo klasifikavimo tikimybės įverčiai, gauti SI metodu.
- *Cv_out* – vardas duomenų žingsnio programos, kurioje saugomos kiekvienos klasės klaidingo klasifikavimo tikimybės įverčiai, gauti KP metodu.
- *Bootstrap_out* – vardas duomenų žingsnio programos, kurioje saugomos kiekvienos klasės klaidingo klasifikavimo tikimybės įverčiai, gauti IK metodu.
- *Mccv_out* – vardas duomenų žingsnio programos, kurioje saugomos kiekvienos klasės klaidingo klasifikavimo tikimybės įverčiai, gauti MKKP metodu.

- *Errors* – vardas duomenų žingsnio programos, kurioje saugomos klaidingo klasifikavimo tikimybės taškiniai ir intervaliniai įverčiai.
- *Median* – vardas duomenų žingsnio programos, kurioje saugomos klaidingo klasifikavimo tikimybės įverčių imčių, gautų IK ir MKKP metodais, medianos ir jų pasikliautinieji intervalai.
- *Class_name* - klasės, kuriai priskiriami klasifikuojami objektai, vardas.

Sukurta vartotojo sąsaja yra patogi, nes vartotojas gali lyginti klasifikavimo metodų taikymo su įvairiomis mokomosiomis imtimis rezultatus. Tam užtenka nurodyti mokomąją imtį, priklausomą bei diskriminavimo kintamuosius. Vartotojo sąsaja yra analogiška SAS vartotojų sukurtoms ir SAS tinklapyje [29] publikuojamoms makrokomandoms. Todėl darbui su autoriaus parašyta makrokomanda pakanka elementarių SAS sistemos žinių. Reikia tik žinoti kas yra duomenų žingsnio programa ir kaip iškviečiama SAS makrokomanda.

2.4 Sukurto įrankio taikymo rekomendacijos ir apribojimai

Prieš kreipiantis į makrokomandą *discrim_compar*, SAS sistemai reikia nurodyti kelią iki failo („discrim_compar.sas“), kuriame saugoma ši bei visos reikalingos pagalbinės makrokomandos. Taip pat reikia nurodyti kelią iki failų („plotit.sas“ ir „multnorm.sas“), kuriuose saugomos SAS makrokomandos *plotit* ir *multnorm*. Tai padaroma naudojant makrokomandą *inc* (kiti metodai pateikti literatūroje [28]):

```
%inc "kelias iki failo\plotit.sas";
%inc "kelias iki failo\multnorm.sas";
%inc "kelias iki failo\discrim_compar.sas";
```

Visų makrokomandoje *discrim_compar* realizuotų veiksmų atlikimui, reikalingos neankstesnės nei 8-tos SAS versijos posistemės: BASE, STAT, GRAPH, IML arba ETS. Su ankstesnėmis versijomis makrokomanda neišbandyta. Makrokomanda gali veikti be GRAPH, IML ir ETS posistemių. Tačiau tuomet nebus atliekami sekantys veiksmai:

- be GRAPH posistemės nebus braižomos taškų sklaidos diagramos;
- be IML ir ETS nebus tikrinama suderinamumo hipotezė, kad diskriminavimo kintamųjų skirstinys yra daugiamatis normalusis. Hipotezė bus tikrinama esant bent viena posistemei IML arba ETS.

Vertinant klaidingo klasifikavimo tikimybę IK bei MKKP metodais, makrokomanda *discrim_compar* sukuria duomenų žingsnio programas (*bootstrap_metodas ir mccv_metodas*), kuriose saugomos atitinkamos IK ir MKKP metodais generuotos klaidingo klasifikavimo tikimybės įverčių imtys.

Kai yra tik dvi klasės, vartotojas gali įvertinti pasirinktų klasifikavimo su mokytoju metodų jautrumą ir specifiškumą. Reikia iš vieneto atimti atitinkamai pirmos ir antros klasės klaidingo klasifikavimo tikimybės įverčius gautus SI metodu. Šie įverčiai pagal nutylėjimą saugomi SAS duomenų žingsnio programoje *resub_errors*.

Vartotojui rekomenduojama dirbti ne standartiniame SAS sistemos *Work* kataloge, o susikurti naują [28], kuriame esantys duomenys nebūtų ištrinti baigus darbą su SAS sistema. Tokiu atveju SAS sitemos darbą nutraukus nenumatytiems trikdžiams, išliktų tarpiniai rezultatai.

2.5 Metodų taikymo rezultatų lyginamoji analizė

Šiame skyrelyje pateikiami sukurto programinio įrankio testavimo bei pasirinktų klasifikavimo su mokytoju metodų taikymo rezultatų lyginamoji analizė. Poskyryje 2.5.1 pateikti pasirinkti klasifikuojami duomenys ir pasirinkimo kriterijai, poskyryje 2.5.2 aprašyti atlikti eksperimentai, o poskyryje 2.5.3 pateikti rezultatai bei jų analizė.

2.5.1 Klasifikuojami duomenys

Darbe pasiūlytos metodikos ir sukurtų programinių priemonių testavimui pasirinkome kelių tipų imtis, kurias sudaro objektai generuojami naudojant trijų skirstinių mišinius (1 lentelė, 4-8 pav.). Imtys skiriasi tuo, kad vienose jų tenkinamos visų klasifikavimo metodų prielaidas, o kitose - tik kelių. Pasirinktos imtys skiriasi ir klasių atskyrimo galimybėmis.

1 lentelė

Naudojamos imtys

A klasės imtys.	B klasės imtys.
$1: n_1 \cdot N((0;0), I)$, $2: n_2 \cdot N((3;2), I)$, $3: n_3 \cdot N((4;0), I)$.	$1: n_1 \cdot N((0;0), I)$, $2: n_2 \cdot N((3;2),5I)$, $3: n_3 \cdot N((4;0),5I)$.
C klasės imtys.	D klasės imtys.
$1: n_1 \cdot N((0;0), I)$, $2: n_{21} \cdot N((3;2),5I) + n_{22} \cdot N((6;4), I)$, $3: n_3 \cdot N((4;0),5I)$.	$1: n_1 \cdot ZG(0,0,30,1,1,1)$, $2: n_2 \cdot ZG(0,0,10,1,1,1)$, $3: n_3 \cdot N((0;0), I)$.
E klasės imtys.	
$1: n_1 \cdot ZG(0,0,5,1,1,1)$, $2: n_2 \cdot ZG(0,0,2.5,1,1,1)$, $3: n_3 \cdot N((0;0), I)$.	

1 lentelėje naudojami žymėjimai: pirmas skaičius žymi klasę, kurios objektai generuojami; antras skaičius (pirmas po dvitaškio) žymi generuojamų objektų skaičių; po jo nurodytas diskriminavimo kintamųjų skirstinys; po pliuso ženklo vėl nurodomas objektų skaičius bei diskriminavimo kintamųjų skirstinys. Pvz., $2:n_{21}\cdot N((3;2),5I)+n_{22}\cdot N((6;4),I)$, - 2-oje klasėje yra n_{21} objektai, kurių diskriminavimo kintamųjų skirstinys dvimatis normalusis $N((3;2),5I)$, ir n_{22} objektai (išskirtys), kurių diskriminavimo kintamųjų skirstinys dvimatis normalusis $N((6;4),I)$.

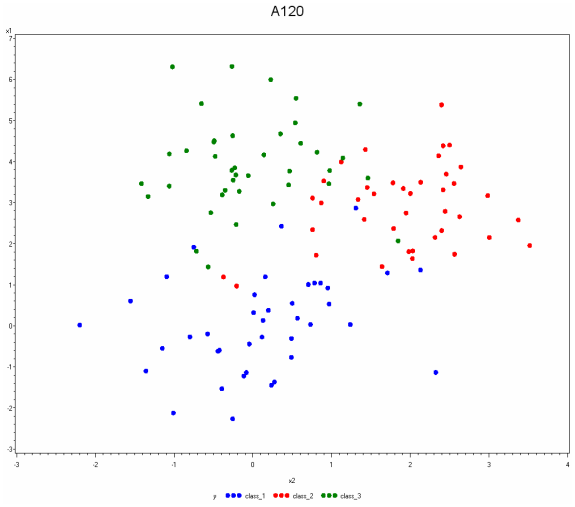
ZG – žiedinis Gauso skirstinys. Atsitiktinį dydį $Y=(X_1,X_2)$ vadiname Žiediniu dvimačiu Gauso dydžiu, jeigu $X_1=\frac{U\cdot Z}{\sqrt{U^2+V^2}}$, $X_2=\frac{V\cdot Z}{\sqrt{U^2+V^2}}$. Čia U, V, Z yra vienmačiai atsitiktiniai dydžiai, $U\sim N(\mu_u,\sigma_u^2)$, $V\sim N(\mu_v,\sigma_v^2)$, $Z\sim N(\mu_z,\sigma_z^2)$. Žymime $Y\sim ZG(\mu_u,\mu_v,\mu_z,\sigma_u^2,\sigma_v^2,\sigma_z^2)$.

A tipo imtis paimta iš populiacijos, kurioje diskriminavimo kintamieji tenkina TDA metodo prielaidas. B tipo imtis paimta iš populiacijos, kurioje diskriminavimo kintamieji tenkina KDA metodo prielaidas. C tipo imtis yra tokia pati kaip ir A tipo tik joje yra išskirčių. D ir E tipo imtys, paimtos iš populiacijų, kuriose diskriminavimo kintamieji netenkina parametrinių klasifikavimo su mokytoju metodų prielaidų.

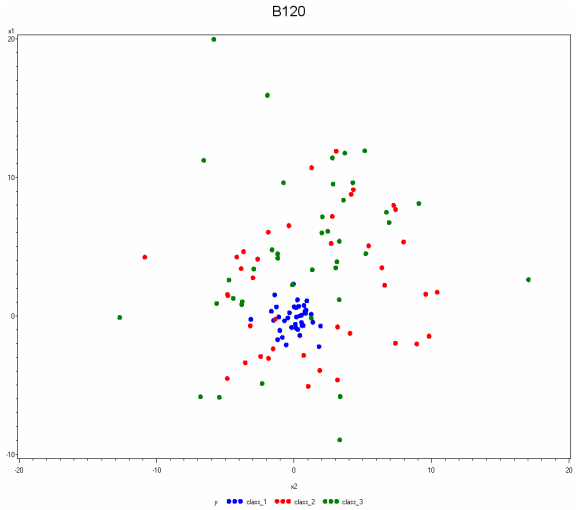
A, B, D ir E imtyse $n_1=n_2=n_3=\frac{n}{3}$, apriorinių tikimybių įverčiai atitinkamai lygūs $\hat{\pi}_1=\hat{\pi}_2=\hat{\pi}_3=\frac{1}{3}$. Klasifikavimui naudosime $n\in\{30,120,300\}$ dydžių A, B, D ir E imtis. C tipo naudosime tik vienį imtį $n=130$, $n_1=40$, $n_{21}=40$, $n_{22}=10$, $n_3=40$. Toliau darbe imtis žymėsime A30, B120, C130 ir t.t., kur raidė parodo imties tipą, o skaičius – imties dydį.

Klasifikavimui pasirinkti duomenys, kuriuose klasifikuojamus objektus apibūdina du diskriminavimo kintamieji, nes šiuo atveju galima pateikti klasifikavimo rezultatus grafikuose.

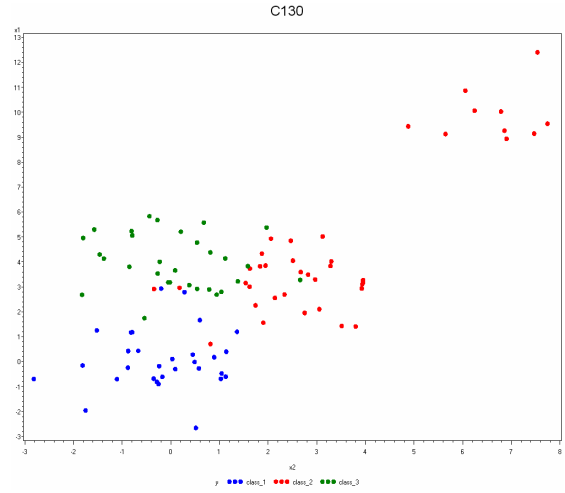
4-8 paveiksluose pateiktos pasirinktų skirtingų imčių tipų objektų sklaidos diagramos. Mėlyna spalva pavaizduoti pirmos klasės objektai, raudona – antros, o žalia – trečios.



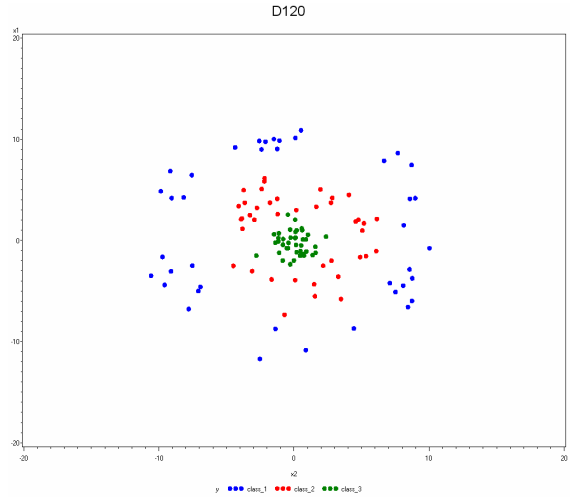
4 pav. Taškų sklaidos diagrama (imtis A120)



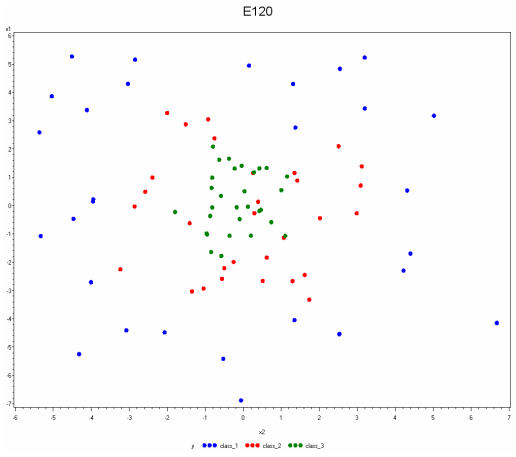
5 pav. Taškų sklaidos diagrama (imtis B120)



6 pav. Taškų sklaidos diagrama (imtis C130)



7 pav. Taškų sklaidos diagrama (imtis D120)



8 pav. Taškų sklaidos diagrama (imtis E120)

Tokių tipų ir dydžių imtys pasirinktos, nes siekiama atsakyti į klausimus:

- kurie klasifikavimo su mokytoju metodai tinkamiausi mažose imtyse (n=30), kurie – vidutinio dydžio imtyse (n=120) ir kurie – didelėse (n=300)?
- Ar skiriasi klasifikavimo rezultatai, gauti TDA ir KDA metodais, kai tenkinamos TDA metodo prielaidos? Ar šis skirtumas statistiškai reikšmingas?
- Ar skiriasi klasifikavimo rezultatai, gauti TDA ir KDA metodais, kai netenkinamos TDA metodo prielaidos? Ar šis skirtumas statistiškai reikšmingas?
- Ar skiriasi klasifikavimo rezultatai, gauti TDA ir LR metodais, kai tenkinamos TDA metodo prielaidos? Ar šis skirtumas statistiškai reikšmingas?
- Ar skiriasi klasifikavimo rezultatai, gauti TDA ir LR metodais, kai netenkinamos TDA metodo prielaidos? Ar šis skirtumas statistiškai reikšmingas?
- Kuriuos metodus (parametrinius ar neparametrinius) rinktis klasifikuojant nagrinėjamų imčių objektus?
- Kurie metodai jautresni išskirtims?

Nagrinėjamos imtys sugeneruotos naudojant autoriaus parašytas makrokomandas *Gauso_class_3* ir *ZG_3* [1 priedas]. Pateikiame kreipinius į makrokomandas, generuojančias imtis A120 ir D120:

```
%Gauso_class_3(
d_out=A120,       m_x1_clas1=0,           m_x1_clas2=3,     m_x1_clas3=4,
x_out=x,          m_x2_clas1=0,           m_x2_clas2=2,     m_x2_clas3=0,
y_out=y,          std_x1_clas1=1,         std_x1_clas2=1,   std_x1_clas3=1,
                  std_x2_clas1=1,         std_x2_clas2=1,   std_x2_clas3=1,
n11=30,
n21=30,           m_x1_clas1_out=0,       m_x1_clas2_out=0  m_x1_clas3_out=0,
n31=30,           m_x2_clas1_out=0,       m_x2_clas2_out=0, m_x2_clas3_out=0,

n12=0,
n22=0,
n32=0 );

%ZG_3   (   d_out=D120,       m_z_clas1=10,           n_clas1=30,
            x_out=x,          m_z_clas2=5,            n_clas2=30,
            y_out=y,                                  n_clas3=30  ).
```

2.5.2 Eksperimentų atlikimo schema

Klaidingo klasifikavimo tikimybes vertinant IK ir MKKP metodais, naudojame 100 iteracijų, nes didinant iteracijų skaičių nuo 100 iki 1000 įverčio tikslumo didėjimas yra minimalus, o laiko sąnaudos išauga beveik 10 kartų [19].

Klaidingo klasifikavimo tikimybes vertinant MKKP metodu, 20% mokomosios imties objektų priskiriame testinei, o likusius – apmokymo imčiai, kaip rekomenduojama literatūroje [18].

Hipotezės tikrinamos su 0,05 reikšmingumo lygmeniu, o intervaliniai įverčiai sudaromi su 0,95 pasikliovimo lygmeniu.

Klasifikuojant imties A30 ir A120 objektus BDA metodu, naudojami penki skirtingi branduoliai. Klasifikuojant kitų imčių objektus naudojamas tik vienas dažniausiai praktikoje taikomas Epanechnikovo branduolys (2.3 formulė).

Taikant BDA ir AKDA metodus, parenkami glodinimo parametrai, su kuriais klaidingo klasifikavimo tikimybės taškinis įvertis KP yra mažiausias. Klaidingo klasifikavimo tikimybės įvertinimo KP metodas pasirinktas dėl dviejų priežasčių:

- jis, panaudojant efektyvų laiko atžvilgiu algoritmą, realizuotas SAS/STAT procedūroje *Discrim* [28].
- Laiko atžvilgiu už kryžminio patikrinimo metodą efektyvesnis yra savos imties metodas. Tačiau šis metodas turi savybę pateikti „optimistinius“ (paslinktus) klaidingo klasifikavimo tikimybės įverčius[21].

Glodinimo parametro parinkimui naudojama autoriaus parašyta makrokomanda *optimal_r* [1 priedas]. Nurodžius pradinę, galutinę glodinimo parametro reikšmes bei žingsnį, kuriuo keičiamas glodinimo parametras, atliekamas klasifikavimas. Pateiksime kreipinius į makrokomandą D120 atveju:

```
%optimal_r  (d_in=D120, start_r=.1, stop_r=1,6, step_r=.1, d_out=BDA_parameter,
             KDA_epa=1);

%optimal_r  (d_in=D120, start_r=1, stop_r=10, step_r=1, d_out=AKDA_parameter,
             NNDA=1).
```

Makrokomanda pateikia gautus klaidingo klasifikavimo tikimybės taškinius įverčius KP skirtingoms glodinimo parametro reikšmėms (2 ir 3 lentelės).

2 lentelė

BDA metodo glodinimo parametrai (imtis D120)

Nr.	Metodas	KP	Glodinimo parametras
1	BDA	0.0833	0.5
2	BDA	0.0833	0.6
3	BDA	0.0917	0.7
4	BDA	0.0917	0.8
5	BDA	0.0917	0.9
6	BDA	0.1167	0.4
7	BDA	0.1250	1.0
8	BDA	0.1333	1.1
9	BDA	0.1750	1.2
10	BDA	0.2000	1.3
11	BDA	0.2083	0.3
12	BDA	0.2417	1.4
13	BDA	0.2917	1.5
14	BDA	0.3250	1.6
15	BDA	0.3500	0.2
16	BDA	0.6333	0.1

3 lentelė

AKDA metodo glodinimo parametrai (imtis D120)

Nr.	Metodas	KP	Glodinimo parametras
1	AKDA	0.0000	4
2	AKDA	0.0167	2
3	AKDA	0.0417	5
4	AKDA	0.0417	6
5	AKDA	0.0500	3
6	AKDA	0.0583	1
7	AKDA	0.0833	8
8	AKDA	0.0917	7
9	AKDA	0.1083	10
10	AKDA	0.1167	9

Parenkami glodinimo parametrai, su kuriais KP yra mažiausias. Taikant AKDA metodą, parenkame nelyginę glodinimo parametro reikšmę, nes tokiu atveju mažiau objektų priskiriama nežinomai klasei (2.5.3 skyrelyje pateiktos objektų priskyrimo nežinomai klasei priežastys). Imtyje D120 AKDA metodu su skirtingais glodinimo parametrais 4, 2 ir 5 sudarėme klasifikavimo taisykles bei klasifikavome 14641 dydžio testinės imties objektus (4, 5, 6 lentelės). Kai glodinimo parametras nelyginis, nežinomai klasei priskirta 0,3% visų objektų. Kai glodinimo parametras yra 2 ir 4, nežinomai klasei priskirta atitinkamai 11,9% ir 7,3%.

4 lentelė

Klasifikavimo AKDA metodu rezultatai, kai glodinimo parametras = 4 (imtis D120)

Number of observations and Percent Classified into y					
	class_1	**class_2**	**class_3**	**Other**	**Total**
Total	9336 63.77	3537 24.16	699 4.77	1069 7.30	14641 100.00
Priors	0.33333	0.33333	0.33333		

5 lentelė

Klasifikavimo AKDA metodu rezultatai, kai glodinimo parametras = 2 (imtis D120)

Number of observations and Percent Classified into y					
	class_1	**class_2**	**class_3**	**Other**	**Total**
Total	9137 62.41	3133 21.40	628 4.29	1743 11.90	14641 100.00
Priors	0.33333	0.33333	0.33333		

6 lentelė

Klasifikavimo AKDA metodu rezultatai, kai glodinimo parametras = 5(imtis D120)

Number of observations and Percent Classified into y					
	class_1	**class_2**	**class_3**	**Other**	**Total**
Total	9698 66.24	4041 27.60	858 5.86	44 0.30	14641 100.00
Priors	0.33333	0.33333	0.33333		

Parinkti glodinimo parametrai (D120 atveju BDA ir AKDA metodų glodinimo parametrai 0,5 ir 5) nurodomi makrokomandai *discrim_compar* (kreipinys pateiktas imties D120 atveju):

```
%discrim_compar (
      d_in=D120,          LDA=1,       resub=1,                pp=0,
      variables=x1 x2,    QDA=1,       cv=1,                   graphs=1,
      class=y,            KDA_biw=0,   bootstrap_iter=100,     box_plot=1,
      class_num=3,        KDA_epa=1,   mccv_iter=100,          intermediate_rez=1,
                          r_epa=0.5    t=20,                   assumption_check=1,
                          KDA_norm=0,                          english=0,
                          KDA_tri=0,                           print=1,
                          KDA_uni=0,
                          NNDA=1,
                          nn=5,
                          LR=1         ).
```

Glodinimo parametrų parinkimas, objektų klasifikavimas ir klasifikavimo rezultatų pateikimas kitoms imtims atliekamas analogiškai D120 atvejui. Atliekant eksperimentus parinkti glodinimo parametrai pateikiami 7 lentelėje.

7 lentelė

Glodinimo parametrai

Imtis	Klasifikavimo su mokytoju metodas	Glodinimo parametras	Imtis	Klasifikavimo su mokytoju metodas	Glodinimo parametras
A30	BDA su dvisvoriu branduoliu	1,9	B120	AKDA	11
	BDA	2,1	B300	BDA	0,8
	BDA su normaliuoju branduoliu	0,7		AKDA	5
	BDA su tolygiuoju branduoliu	2,0	C130	BDA	1,6
	BDA su trisvoriu branduoliu	1,9		AKDA	7
	AKDA	3	D30	BDA	1,2
A120	BDA su dvisvoriu branduoliu	1,9		AKDA	3
	BDA	2,5	D120	BDA	0,5
	BDA su normaliuoju branduoliu	0,7		AKDA	5
	BDA su tolygiuoju branduoliu	1,5	D300	BDA	0,5
	BDA su trisvoriu branduoliu	1,9		AKDA	5
	AKDA	5	E30	BDA	1,3
A300	BDA	1,3		AKDA	3
	AKDA	7	E120	BDA	0,8
B30	BDA	2		AKDA	3
	AKDA	5	E300	BDA	0,6
B120	BDA	0,6		AKDA	7

BDA – branduolinės diskriminantinės analizės metodas su Epanechnikovo branduoliu.

2.5.3 Rezultatai ir jų analizė

Šiame skyrelyje pateikta detali vieno eksperimento atlikto su imtimi D120 rezultatų analizė ir visų atliktų eksperimentų pagrindinės išvados. Su kitomis imtimis gauti rezultatai pateikti 2 ir 3 prieduose. Jų analizė atlikta analogiškai imties D120 atvejui.

Eksperimento su imtimi D120 rezultatai ir jų analizė. Pateikiant rezultatus skliausteliuose nurodomas parametras ir jo reikšmė, kuriai esant tie rezultatai gauti.

Jeigu vartotojas makrokomandoje *dirsim_compar* nenurodo išvesti tarpinių rezultatų (*intermediate_rez=0*), tuomet pateikiamos dvi pagrindinių rezultatų lentelės su klasifikavimo metodų kokybę apibūdinančiomis statistikomis (8 ir 9 lentelės). Pirmoje pagrindinėje lentelėje (8 lentelė) pateikiamos šios statistikos:

- SI – klaidingo klasifikavimo tikimybės taškinis įvertis, gautas SI metodu.

- KP – klaidingo klasifikavimo tikimybės taškinis įvertis, gautas KP metodu.
- IK – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos IK metodu, vidurkis.
- MKKP – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos MKKP metodu, vidurkis.
- (IK_95a; IK_95v) – klaidingo klasifikavimo tikimybės 95% pasikliautinasis intervalas, gautas IK metodu.

8 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D120)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0000	0.0417	0.0971	0.0000	0.2073	0.0626
2	BDA	0.0167	0.0833	0.1333	0.0367	0.2386	0.0926
3	KDA	0.0417	0.0583	0.0966	0.0000	0.2142	0.0596
4	LR	0.5417	0.6917	0.6688	0.4167	0.8750	0.6750
5	TDA	0.5417	0.6000	0.5919	0.4671	0.7259	0.6029

AKDA metodo SI ir KP rodo, kad imtyje D120 AKDA metodas yra tinkamiausias. Tačiau IK ir MKKP rodo, kad šioje imtyje tinkamiausias KDA metodas. Didžiausi klaidingo klasifikavimo tikimybės taškiniai įverčiai ($0.541 < \hat{R} \le 0.675$) gaunami klasifikuojant LR ir TDA metodais. LR ir TDA metodų SI, KP, IK ir MKKP didesni nei AKDA, BDA ir KDA metodų mažiausiai 4,4 karto. TDA metodo klaidingo klasifikavimo tikimybės taškiniai įverčiai didesni nei LR metodo. Taigi, naudodami SI, KP, IK ir MKKP, klasifikavimo su mokytoju metodus pagal klasifikavimo kokybę imtyje D120 išrikiuojame tokia tvarka: 1 vieta – AKDA, BDA ir KDA, 2 vieta – TDA, 3 vieta – LR.

Tačiau, naudojant taškinius įverčius negalime teigti, kad vienas ar kitas klasifikavimo su mokytoju metodas statistiškai reikšmingai geresnis ar blogesnis. Tam naudojamas klaidingo klasifikavimo tikimybės 95% pasikliautinasis intervalas (IK_95a; IK_95v), gautas įkelčių metodu. Nagrinėjamu atveju gavome, kad klasifikavimo su mokytoju metodai pagal klasifikavimo kokybę rikiuojami tokia tvarka: 1 vieta – AKDA, BDA ir KDA, 2 vieta – TDA ir LR. AKDA, BDA ir KDA metodų klaidingo klasifikavimo tikimybės intervaliniai įverčiai patenka į intervalą nuo 0 iki 0.24, o TDA ir LR – į intervalą $(0.41; 0.88)$. AKDA, BDA ir KDA metodų klaidingo klasifikavimo tikimybės pasikliautinieji intervalai persidengia, todėl negalime teigti, kad šie metodai skiriasi statistiškai reikšmingai pagal klasifikavimo kokybę Analogiški rezultatai gauti taikant TDA ir LR metodus.

Be SI, KP, IK, MKKP ir (IK_95a, IK_95v) naudojamos dar ir tokios klasifikavimo su mokytoju metodų kokybę apibūdinančios statistikos (9 lentelė):

- IK_me – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos IK metodu, mediana;
- MKKP_me – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos MKKP metodu, mediana;
- (IK_me_95a; IK_me_95v) – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos IK metodu, medianos 95% pasikliautinasis intervalas;
- (MKKP_me_95a; MKKP_me_95v) – klaidingo klasifikavimo tikimybės taškinių įverčių imties, gautos MKKP metodu, medianos 95% pasikliautinasis intervalas.

9 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D120)

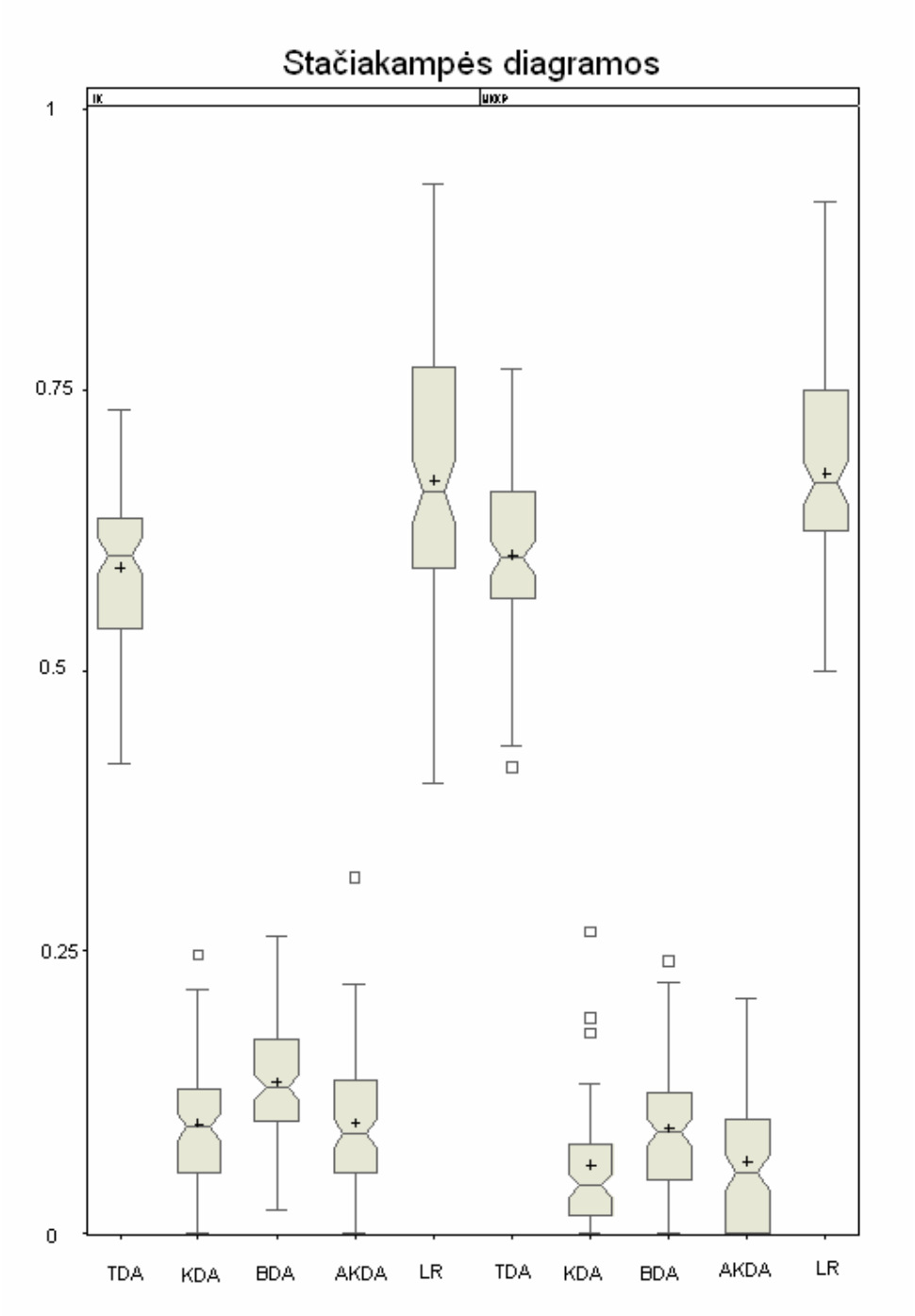
Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.0877	0.0769	0.1056	0.0541	0.0370	0.0667
2	BDA	0.1284	0.1167	0.1396	0.0893	0.0741	0.1037
3	KDA	0.0938	0.0818	0.1085	0.0417	0.0417	0.0667
4	LR	0.6593	0.6279	0.7179	0.6667	0.6250	0.6667
5	TDA	0.6019	0.5815	0.6227	0.6004	0.5852	0.6143

Mažiausi IK_me ir MKKP_me yra AKDA ir KDA metodų. BDA metodo IK_me ir MKKP_me už AKDA ir KDA metodų įverčius didesni daugiau nei 35%, LR ir TDA metodų – daugiau nei 6 kartus. Tačiau naudojant IK_me ir MKKP_me negalime teigti, kad klasifikavimo su mokytoju taikymo rezultatų kokybė statistiškai reikšmingai skiriasi.

Naudodami (IK_me_95a; IK_me_95v) ir (MKKP_me_95a; MKKP_me_95v) klasifikavimo su mokytoju metodus pagal klasifikavimo kokybę išrikiuojame taip:1 vieta – AKDA ir KDA, 2 vieta – BDA, 3 vieta – TDA, 4 vieta – LR. AKDA ir KDA metodų (IK_me_95a; IK_me_95v) ir (MKKP_me_95a; MKKP_me_95v) persidengia, todėl negalime, teigti, kad šių metodų kokybė statistiškai reikšmingai skiriasi.

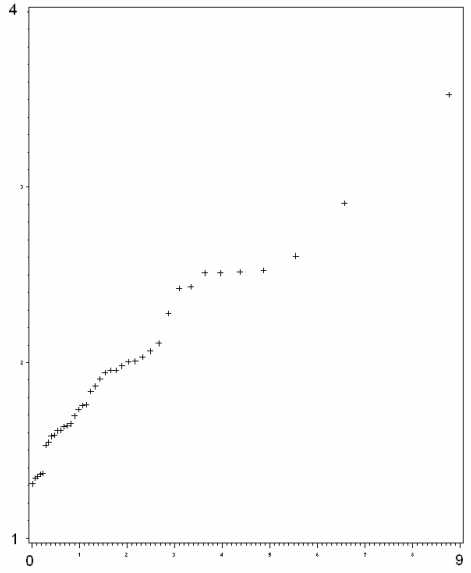
Vartotojui nurodžius (*box_plot=1*), pateikiamas klaidingo klasifikavimo tikimybės taškinių įverčių imčių, gautų IK ir MKKP metodais, stačiakampių diagramų su išpjovomis grafikas (9 pav.) Grafike pirmos penkios yra TDA, KDA, BDA, AKDA ir LR metodų klaidingo klasifikavimo tikimybės taškinių įverčių imčių, gautų IK metodu, stačiakampės diagramos, o sekančios penkios - TDA, KDA, BDA, AKDA ir LR metodų klaidingo klasifikavimo tikimybės taškinių įverčių imčių, gautų MKKP metodu, stačiakampės diagramos. Stačiakampėse diagramose naudojami tokie žymėjimai: horizontalus brūkšnys stačiakampyje – IK_me arba MKKP_me, pliusas – IK arba MKKP,

stačiakampio kraštinės ilgis – kvartilinis plotis *IQR* (lygūs 3-čio ir 1-o kvartilių skirtumui, angl. *interquartile range*), išpjovos stačiakampyje - (IK_me_95a; IK_me_95v) arba (MKKP_me_95a; MKKP_me_95v), kvadratėliai žymi išskirtis, t.y. objektus nuo pirmo ir antro kvartilių nutolusius didesniu nei $1.5 \cdot IQR$ atstumu [28].

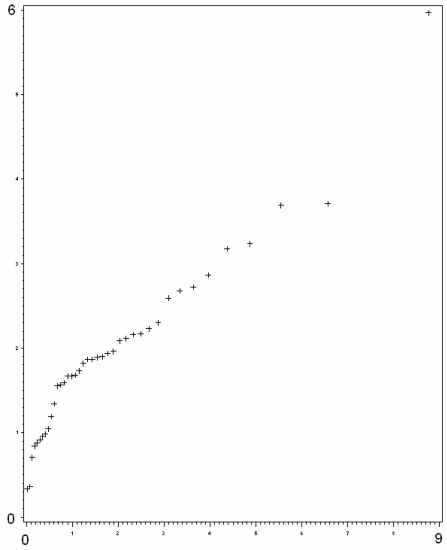


9 pav. Stačiakampės diagramos (imtis D120)

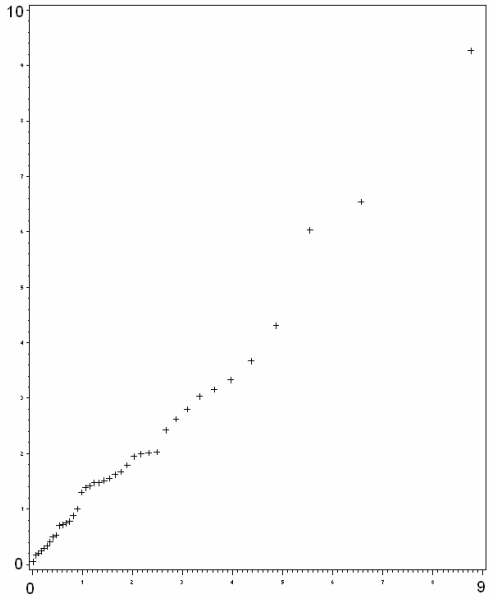
Prieš taikant parametrinius klasifikavimo su mokytoju metodus, reikia patikrinti ar tenkinamos metodo prielaidos (*asumption_check=1*). Apie normalumo sąlygos tenkinimą grubiai galime spręsti iš Mahalanobio atstumų kvadratų „Q-Q“ grafikų.



10 pav.Mahalanobio atstumų kvadratų „Q-Q“ grafikas (imtis D120, 1-klasė)



11 pav.Mahalanobio atstumų kvadratų „Q-Q“ grafikas(imtis D120, 2-klasė)



12 pav.Mahalanobio atstumų kvadratų „Q-Q“ grafikas (imtis D120, 3-klasė)

Pirmos ir antros klasių grafikų (10, 11 pav.) taškai nukrypę nuo tiesės y=x, tai rodo, kad diskriminavimo kintamųjų skirstinys nėra daugiamatis normalusis. Trečios klasės grafiko (12 pav.) taškai išsibarstę apie tiesę y=x, todėl diskriminavimo kintamųjų skirstinys šioje klasėje gali būti

daugiamatis normalusis. Mahalanobio atstumų kvadratų „Q-Q“ grafikai yra grubus ir subjektyvus daugiamačio normalumo tikrinimo testas. Jis daugiau skirtas nustatyti, ar imtyje nėra išskirčių (nagrinėjamu atveju jų nėra), dėl kurių gali būti atmetamos suderinamumo hipotezės.

10 lentelė

Normalumo tikrinimas (imtis D120, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.90	0.0018
x2	Shapiro-Wilk W	0.87	<.0001
System	Mardia Skewness	0.73	0.9478
	Mardia Kurtosis	-2.98	0.0029
	Henze-Zirkler T	4.53	<.0001

11 lentelė

Normalumo tikrinimas (imtis D120, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.92	0.0069
x2	Shapiro-Wilk W	0.90	0.0024
System	Mardia Skewness	3.81	0.4316
	Mardia Kurtosis	-2.27	0.0230
	Henze-Zirkler T	3.78	0.0002

12 lentelė

Normalumo tikrinimas (imtis D120, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.6872
x2	Shapiro-Wilk W	0.99	0.9711
System	Mardia Skewness	2.38	0.6655
	Mardia Kurtosis	-0.25	0.8058
	Henze-Zirkler T	-1.05	0.2931

13 lentelė

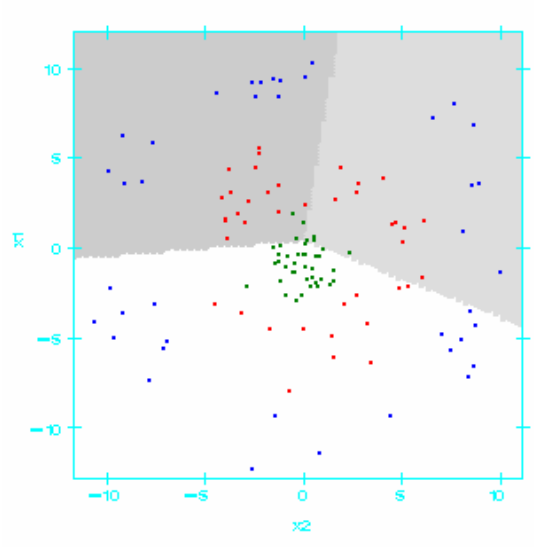
Kovariacijų matricų homogeniškumo tikrinimas (imtis D120)

Chi-Square	DF	Pr > ChiSq
201.819142	6	<.0001

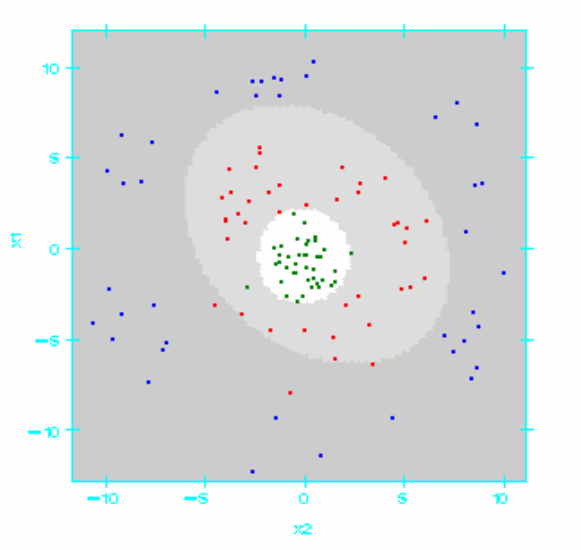
10 11 ir 12 lentelėse pateikiami kiekvienos klasės (1, 2 ir 3) suderinamumo hipotezių tikrinimo rezultatai, t.y. naudojamų statistikų reikšmės bei *p*-reikšmės. Imties D120 1-oje ir 2-oje klasėje visų suderinamumo hipotezei naudojamų kriterijų, išskyrus Mardijos asimetrijos koeficiento kriterijų, *p*-reikšmės mažesnės už reikšmingumo lygmenį (0.05). Vadinasi, hipotezė, kad diskriminavimo kintamųjų skirstinys imties D120 1-oje ir 2-oje klasėje yra daugiamatis normalusis, atmetama. Todėl parametrinių klasifikavimo su mokytoju metodų TDA ir KDA taikymas, klasifikuojant D120 imties objektus, yra nekorektiškas.

Kovariacijų matricų homogeniškumo hipotezė (13 lentelė) taip pat atmesta ($p<0.0001$). Todėl, net jei ir būtų tenkinama diskriminavimo kintamųjų normalumo sąlyga, TDA metodo taikymas būtų nekorektiškas. Tačiau tokiu atveju korektiškai galėtume taikyti KDA metodą.

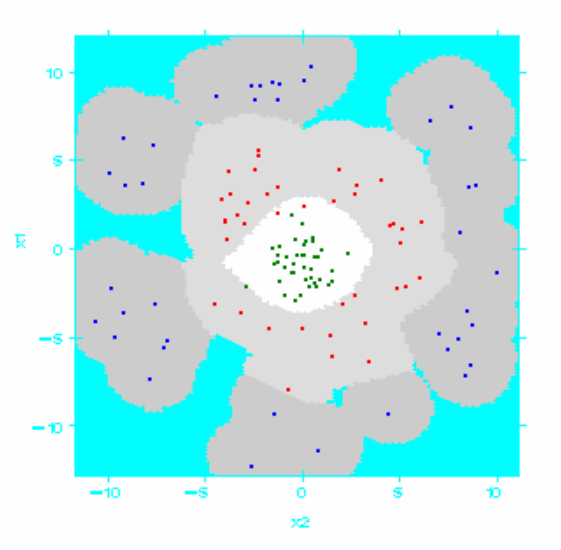
Renkantis klasifikavimo su mokytoju metodą, naudinga žinoti kokios kreivės ar paviršiai atskiria klases. Vartotojui nurodžius (*graph=1)*, pateikiami imčių klasifikavimo rezultatų grafikai, kuriuose tamsiausia spalva nuspalvinta 1 klasė, šviesesne – 2 klasė, balta – 3 klasė, o melsva – nežinoma klasė. Nežinoma klasę sudaro objektai, kurie naudojantis sudarytomis klasifikavimo taisyklėmis nepriskirti nei vienai iš anksto žinomai klasei.



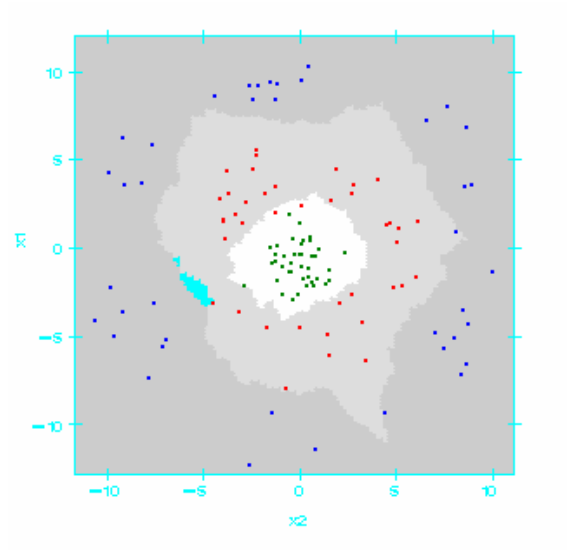
13 pav.Klasifikavimo TDA metodu rezultatai (imtis D120)



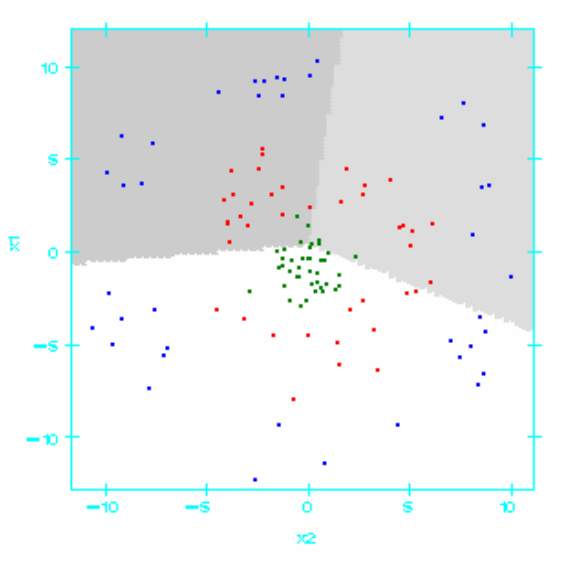
14 pav.Klasifikavimo KDA metodu rezultatai (imtis D120)



15 pav.Klasifikavimo BDA metodu rezultatai (imtis D120)



16 pav.Klasifikavimo AKDA metodu rezultatai (imtis D120)



17 pav.Klasifikavimo LR metodu rezultatai (imtis D120)

Taikant TDA ir LR metodus, objektų klasės atskiriamos tiesėmis (13 ir 17 pav.), o KDA metodą – elipsėmis (14 pav.) (bendru atveju – antros eilės kreivėmis). Taikant BDA ir AKDA metodus klases skiriančių kreivių formos (15 ir 16 pav.) daug sudėtingesnės.

Klasifikuojant objektus BDA ir AKDA metodais, tam tikri objektai priskiriami nežinomai klasei (paveiksluose pavaizduota melsva spalva). BDA metode tokių objektų yra daugiau nei AKDA (atitinkamai 23% ir 0.3%, 19 ir 6 lentelės). BDA metode objektas priskiriamas nežinomai klasei, kai lokalioje objekto aplinkoje (jos dydis priklauso nuo glodinimo parametro) nėra nei vieno apmokymo imties objekto. Tokiu atveju kiekvienos klasės tankio įvertis lygus nuliui ir negalime rasti aposteriorinės tikimybės įverčio.

Taikant AKDA metodą objektas priskiriamas nežinomai klasei, jeigu lokalioje objekto aplinkoje yra vienodas skirtingoms klasėms priklausančių objektų skaičius (tokiu atveju gauname, kad tų klasių aposteriorinių tikimybių įverčiai lygūs).

Autoriaus sukurta metodika bei ją įgyvendinantis programinis įrankis suteikia vartotojui galimybę (*intermediate_rez=1*) lyginti kiekvienos klasės SI, KP, IK ir MKKP (14-17 lentelės). Jie gaunami analogiškai kaip ir atitinkami bendros klaidingo klasifikavimo tikimybės taškiniai įverčiai.

14 lentelė

Klasių SI įverčiai (imtis D120)

Nr.	Metodas	class_1	class_2	class_3
1	AKDA	0.0000	0.0000	0.0000
2	BDA	0.0000	0.0500	0.0000
3	KDA	0.0000	0.0500	0.0750
4	LR	0.6500	0.6750	0.3000
5	TDA	0.6500	0.6750	0.3000

15 lentelė

Klasių KP įverčiai (imtis D120)

Nr.	Metodas	class_1	class_2	class_3
1	AKDA	0.0250	0.1000	0.0000
2	BDA	0.1500	0.1000	0.0000
3	KDA	0.0000	0.0750	0.1000
4	LR	0.8000	0.7750	0.5000
5	TDA	0.7750	0.7250	0.3000

16 lentelė

Klasių IK įverčiai (imtis D120)

Nr.	Metodas	class_1	class_2	class_3
1	AKDA	0.1132	0.1608	0.0173
2	BDA	0.2360	0.1492	0.0149
3	KDA	0.0408	0.1539	0.0952
4	LR	0.7014	0.7065	0.5179
5	TDA	0.6970	0.7281	0.3507

17 lentelė

Klasių MKKP įverčiai (imtis D120)

Nr.	Metodas	class_1	class_2	class_3
1	AKDA	0.0826	0.1035	0.0018
2	BDA	0.1776	0.0900	0.0102
3	KDA	0.0059	0.0902	0.0826
4	LR	0.7392	0.7091	0.4530
5	TDA	0.7682	0.7156	0.3250

AKDA BDA ir KDA klasifikavimo su mokytoju metodų jautrumai lygūs 1 (jautrumas = 1 – pirmos klasės SI), o LR ir TDA metodų jautrumai – 0.35. Pagal jautrumą, kaip ir SI, klasifikavimo su apmokymu metodai pagal 1-klasės objektų klasifikavimo kokybę rikiuojami tokia tvarka: 1 vieta – AKDA, BDA ir KDA, 2 vieta – LR ir TDA. Kadangi vertinant jautrumą naudojamas 1-klasės SI, tai jautrumo matas, kaip ir SI, yra „optimistinis“. Jeigu svarbu teisingiau klasifikuoti pirmos klasės objektus, reiktų, pasirenkant klasifikavimo metodą, naudoti pirmos klasės ne tik SI, bet ir KP, IK bei MKKP. Jie parodo, kad geriausiai pirmos klasės objektus klasifikuoja KDA metodas ($\hat{R} < 0.05$), antras AKDA metodas ($\hat{R} \approx 0.1$), trečias – BDA ($\hat{R} \approx 0.2$), o blogiausi klasifikatoriai – LR ir TDA ($\hat{R} > 0.69$). Pirmos klasės KP klasifikavimo metodus pagal klasifikavimo kokybę išrikiuoja taip pat kaip IK bei MKKP, tik skiriasi klaidingo klasifikavimo tikimybės dydžiai: KDA metodo $\hat{R} = 0$, o AKDA - $\hat{R} = 0.025$. Kadangi klasių yra daugiau nei dvi, tai specifiškumas šiuo atveju neskaičiuojamas.

Eksperimento su imtimi D120 išvados. Atlikus imties D120 klasifikavimo rezultatų analizę, galime teigti, kad geriausiai (klaidingo klasifikavimo tikimybės įverčių, gautų IK ir MKKP metodais, imties medianos 95% pasikliautinojo intervalo prasme) imties D120 objektus klasifikuoja AKDA ir KDA metodai. Kadangi netenkinamos parametrinio KDA metodo taikymo prielaidos, tai klasifikuojant imties D120 objektus reiktų rinktis AKDA metodą. Šios imties objektams klasifikuoti visai netinkami LR ir TDA metodai.

Eksperimentų, atliktų su A, B, C, D, E tipo imtimis, rezultatų analizė. Imties A30 atveju klasifikavimo kokybe išsiskiria AKDA, TDA ir BDA su normaliuoju branduoliu metodai. Jų SI ir MKKP vidutiniškai 2 kartus mažesni nei kitų klasifikavimo su mokytoju metodų (20 lentelė). AKDA, TDA ir BDA su normaliuoju branduoliu klaidingo klasifikavimo tikimybės taškiniai įverčiai tarpusavyje skiriasi labai nedaug (IK atitinkamai lygūs 0.077, 0.088, 0.059). Kadangi tenkinamos visos parametrinio TDA metodo prielaidos (44-47 lentelės), tai klasifikuojant A30 imties objektus reiktų rinktis paprasčiausią TDA metodą. Iš branduolinės diskriminantinės analizės metodų didžiausia klasifikavimo kokybe pasižymi BDA su normaliuoju branduoliu (jo KP, IK ir MKKP vidutiniškai 2 kartus mažesni nei kitų).

Klasifikuojant imties A120 objektus reiktų rinktis paprasčiausią parametrinį TDA metodą, kurio SI=KP=IK=0.12 (22 lentelė) ir kurio taikymas šioje imtyje yra korektiškas, nes tenkinamos visos prielaidos (48-51 lentelės). Prasčiausiai A120 imties objektus klasifikuoja AKDA metodas, kurio KP=MKKP=0.15. AKDA metodo SI=0.0917, o tai patvirtina, kad AKDA metodas prisitaiko prie apmokymo imties ir kad SI yra „optimistinis“ paslinktas įvertis. BDA su Epanechnikovo branduoliu KP, IK ir MKKP mažesni nei BDA su kitais branduoliais KP, IK ir MKKP (22 lentelė). Tai patvirtina, kad skirtingose imtyse tinkamesnė (SI, KP, IK, MKKP prasme) gali būti vis kita branduolio funkcija (imtyje A30 tinkamiausia normalioji, o imtyje A120 - Epanechnikovo).

Klasifikuojant imties A300, kaip ir visų A tipo imčių, objektus tinkamiausias parametrinis TDA metodas, kurio taikymas šioje imtyje yra korektiškas (52-54 lentelės). Blogiausiai imties A300 objektus klasifikuoja AKDA metodas. Šio metodo klasifikavimo kokybė statistiškai reikšmingai skiriasi nuo TDA metodo klasifikavimo kokybės pagal (IK_me_95a; IK_me_95v) (AKDA metodo - (0.163; 0.172), o TDA – (0.131; 0.144)) (25 lentelė).

Klasifikuojant B tipo imčių objektus tinkamiausias yra parametrinis KDA metodas, kurio taikymas yra korektiškas (56-58, 60-62, 64-66 lentelės). Visose B tipo imtyse TDA metodo taikymas nekorektiškas, nes netenkinama kovariacijų matricų homogeniškumo sąlyga (59, 63 ir 67 lentelės). B tipo imtyse KDA metodas pagal klasifikavimo kokybę statistiškai reikšmingai išsiskiria iš kitų metodų (KDA metodo (MKKP_me_95a; MKKP_me_95v) imtyse B30, B120 ir B300 atitinkamai (0.278; 0.333), (0.347; 0.380), (0.345; 0.367)). Pagal (MKKP_me_95a; MKKP_me_95v) statistiškai reikšmingai blogiausi B tipo imtyse yra TDA ir LR metodai.

Imtyje C130, lyginant su A120, TDA ir KDA metodų SI ir KP padidėja (daugiau nei 10%), o AKDA, BDA ir LR metodų SI ir KP sumažėja (vidutiniškai 17%) (22 ir 32 lentelės). Vadinasi TDA ir KDA klasifikavimo su mokytoju metodai jautrūs išskirtims. Klasifikuojant imties C130 objektus reiktų rinktis neparametrinį BDA metodą. Nedaug klasifikavimo kokybe nuo BDA skiriasi paprastesnis LR metodas (jų IK atitinkamai 0.129 ir 0.134, o pasikliautinieji intervalai persidengia) (32, 33 lentelės). Parametrinių TDA ir KDA metodų taikymas, klasifikuojant imties C130 objektus yra nekorektiškas, nes suderinamumo hipotezės, kad 2-oje klasėje diskriminavimo kintamųjų skirstiniai yra vienmačiai normalieji, atmestos (69 lentelė). Taip pat atmesta ir kovariacijų matricų homogeniškumo hipotezė $p<0.0001$ (71 lentelė).

Imtyse D30 ir D300 klasifikavimo kokybe išsiskiria neparametriniai AKDA ir BDA metodai. AKDA ir BDA metodų SI, KP, IK ir MKKP imtyje D30 vidutiniškai daugiau nei 2 kartus mažesni nei TDA ir LR metodų, o imtyje D300 – daugiau nei 16 kartų (34, 36 lentelės). AKDA, BDA metodai statistiškai reikšmingai kokybiškiau, pagal (IK_me_95a; IK_me_95v) ir (MKKP_me_95a; MKKP_me_95v), klasifikuoja imčių D30 ir D300 objektus (imtyje D300 AKDA ir BDA metodų $MKKP\_me\_95v < 0.016$, o kitų klasifikavimo su mokytoju metodų $MKKP\_me\_95a > 0.017$) (35, 37 lentelės). AKDA ir BDA metodų tarpusavyje klasifikavimo kokybė statistiškai reikšmingai nesiskiria. Imtyje D30 BDA metodo SI, KP, IK ir MKKP vidutiniškai 14% mažesni nei AKDA metodo (34 lentelė), o imtyje - vidutiniškai 38% didesni (36 lentelė). Todėl klasifikuojant imties D30 objektus reiktų taikyti BDA metodą, o klasifikuojant imties D300 objektus – AKDA metodą. Statistiškai reikšmingai blogiausiai, pagal (IK_ 95a; IK_95v), imties D300 objektus klasifikuoja TDA ir LR metodai, jų $IK\_95a > 0.45$, o kitų klasifikavimo su mokytoju metodų $IK\_95v < 0.1$ (36 lentelė).

Naudodami klaidingo klasifikavimo tikimybės 95% pasikliautinąjį intervalą, apskaičiuotą IK metodu, imtyje E30 statistiškai reikšmingai negalime išskirti nei vieno metodo (38 lentelė). Imtyje E120 KDA ir BDA metodai ($IK\_95v < 0.42$) statistiškai reikšmingai geresni už LR ir TDA ($IK\_95a > 0.45$) (40 lentelė). Imtyje E300 KDA, BDA ir AKDA metodai ($IK\_95v < 0.35$) statistiškai reikšmingai geresni už LR ir TDA ($IK\_95a > 0.53$) (42 lentelė). E tipo imtyse pagal klasifikavimo kokybę, naudodami KP, IK ir MKKP, klasifikavimo su mokytoju metodus surikiuojame taip: 1-as – KDA, 2-as – BDA, 3-ias – AKDA, 4-as – TDA, 5-as – LR (38, 40 ir 42 lentelės). KDA metodo prielaidos tenkinamos tik imtyje E30 (80-83 lentelės), o kitose E tipo imtyse netenkinamos (84-91 lentelės). Vadinasi klasifikuojant E30, E120 ir E300 objektus, reiktų atitinkamai rinktis KDA, BDA, BDA metodus.

Atlikus eksperimentų rezultatų analizę, nustatyti klasifikavimo su mokytoju metodai, tinkamiausi analizuotų imčių objektų klasifikavimui (18 lentelė).

18 lentelė

Tinkamiausi klasifikavimo su mokytoju metodai

Imtis	Tinkamiausias metodas	SI	KP	IK	MKKP
A30	TDA	0.0667	0.0667	0.0878	0.0874
A120	TDA	0.1167	0.1167	0.1185	0.1071
A300	TDA	0.1267	0.1300	0.1393	0.1265
B30	KDA	0.1667	0.3333	0.4033	0.3328
B120	KDA	0.3250	0.4000	0.3642	0.3657
B300	KDA	0.3400	0.3600	0.3623	0.3551
C130	BDA	0.0833	0.0944	0.1288	0.1052
D30	BDA	0.0667	0.1000	0.2558	0.1643
D120	AKDA	0.0000	0.0417	0.0971	0.0626
D300	AKDA	0.0033	0.0067	0.0202	0.0091
E30	KDA	0.2333	0.3333	0.3844	0.3486
E120	BDA	0.1222	0.1778	0.2671	0.2107
E300	BDA	0.1767	0.2167	0.2529	0.2244

TDA ir KDA metodų taikymo rezultatų lyginimas. TDA metodas statiškai reikšmingai geriau klasifikuoja tik imties A30 objektus (TDA metodo IK_me_95v = 0.111, o KDA metodo IK_me_95a = 0.133) (21 lentelė). Imtyse A120 ir A300 TDA metodas tinkamesnis, tačiau skirtumai nėra statistiškai reikšmingi (22-25 lentelės). B, D ir E tipo imtyse statistiškai reikšmingai, pagal (IK_me_95a; IK_me_95v) ir (MKKP_me_95a; MKKP_me_95v), tinkamesnis KDA metodas (8, 9, 26-31, 34-43 lentelės). Imtyje C130 KDA metodas geresnis nei TDA, tačiau skirtumas statistiškai nereikšmingas (32, 33 lentelės). Vadinasi, kai tenkinamos TDA metodo prielaidos reikia rinktis TDA metodą, priešingu atveju - KDA.

TDA ir LR metodų taikymo rezultatų lyginimas. TDA ir LR metodų taikymo rezultatus lyginant pagal (IK_me_95a; IK_me_95v), imtyse A30, B ir D tipo, E30, E300 statistiškai reikšmingai geresnis TDA metodas (21, 27, 29, 31, 35, 37, 39, 41, 45 lentelės). Imtyse A120, A300, E120 TDA metodas geresnis, tačiau skirtumai statistiškai nereikšmingi (23, 25, 9 lentelės). Imtyje C130 LR metodo SI, KP, IK ir MKKP vidutiniškai 10% didesni nei TDA metodo (32 lentelė), tačiau šis skirtumas statistiškai nereikšmingas (33 lentelė). Todėl, kai tenkinamos TDA metodo prielaidos, renkantis tarp TDA ir LR pirmenybę reikia teikti TDA metodui. Eksperimentų rezultatai parodė, kad TDA metodo klasifikavimo

kokybė geresnė už LR net tose imtyse (B, D tipo), kur netenkinamos TDA metodo prielaidos. Kai duomenyse yra išskirčių, reiktų rinktis LR metodą.

BDA ir AKDA metodų taikymo rezultatų lyginimas. Imtyse A120, A300, B120, D120, D300 statistiškai reikšmingai pagal (IK_me_95a; IK_me_95v) geresnis AKDA metodas (23, 25, 29, 9, 37 lentelės), o imtyse E30, E120 – BDA metodas (39, 41 lentelės). Imtyse A30, B30, B300, C130, D30, E300 AKDA ir BDA metodų klasifikavimo kokybė statistiškai reikšmingai nesiskiria (21, 27, 31, 33, 35, 43 lentelės). Kadangi imčių, kuriose tinkamesnis vienas metodas, nesieja joks bendras požymis (pvz., imties dydis, tipas), tai negalime nustatyti kriterijų ir taisyklių kada kurį metodą pasirinkti. Abiejų neparametrinių metodų taikymo rezultatai labai priklauso nuo parinktų glodinimo parametrų, o BDA metodas dar ir nuo parinkto branduolio (4-6, 20-23 lentelės).

Parametrinių ir neparametrinių metodų taikymo rezultatų lyginimas. Kai tenkinamos parametrinių klasifikavimo su mokytoju metodų taikymo prielaidos, reikia rinktis parametrinius metodus (eksperimentų su A ir B tipo imtimis rezultatai). Neparametrinių BDA ir AKDA metodų trūkumas lyginant juos su parametriniais yra tas, kad jų klasifikavimo kokybė priklauso nuo parinkto glodinimo parametro. Jei parenkam tinkamą glodinimo parametrą, tai BDA ir AKDA metodų klasifikavimo kokybė statistiškai reikšmingai nesiskiria nuo parametrinių (imties A30 atveju TDA metodo klasifikavimo kokybė statistiškai reikšmingai nesiskiria nuo BDA ir AKDA (21 lentelė), o B30 atveju KDA nesiskiria nuo BDA ir AKDA (27 lentelė)). Tačiau parinkus netinkamą glodinimo parametrą, neparametrinių metodų kokybė statistiškai reikšmingai skiriasi nuo parametrinių (imties A300 atveju TDA geresnis už AKDA (25 lentelė), o B120 atveju KDA geresnis už BDA (29 lentelė)). Kitas BDA ir AKDA metodų trūkumas yra tas, kad juos taikant reikia atlikti daugiau skaičiavimų nei taikant parametrinius TDA ir KDA. Kai parametrinių metodų taikymo prielaidos netenkinamos, reikia rinktis neparametrinius metodus.

Klaidingo klasifikavimo tikimybės įverčių analizė. Atlikti eksperimentai patvirtina, kad klaidingo klasifikavimo tikimybės taškiniai įverčiai, gauti savos imties metodu, yra optimistiški. Analizuotų klasifikavimo su mokytoju metodų taikymo pasirinktose imtyse rezultatai parodė, kad klaidingo klasifikavimo tikimybės taškiniai įverčiai, gauti kryžminio patikrinimo, įkelčių bei Monte Karlo kryžminio patikrinimo metodais, vidutiniškai 69% didesni nei gauti savos imties metodu (20, 22,..., 42 lentelės).

Naudojant klaidingo klasifikavimo tikimybės 95% pasikliautinuosius intervalus, nei vienoje imtyje nepavyko išskirti kažkurio vieno klasifikavimo su mokytoju metodo. Pagal šiuos klaidingo klasifikavimo tikimybės intervalinius įverčius tik keturiose imtyse pavyko klasifikavimo metodus padalinti į dvi grupes: imtyse D120, D300, E300 AKDA, BDA ir KDA metodų klasifikavimo kokybė statistiškai reikšmingai geresnė nei TDA ir LR (8, 36, 42 lentelės), o imtyje E120 BDA ir KDA metodų - nei TDA ir LR (40 lentelė). Kai metodai pagal klasifikavimo kokybę atskiriami naudojant

klaidingo klasifikavimo tikimybės 95% pasikliautinuosius intervalus, tai tų metodų klaidingo klasifikavimo tikimybės taškiniai įverčiai skiriasi kelis ar net keliolika kartų (imtyje D120 – vidutiniškai 13 kartų, imtyje D300 – 49, imtyje E300 – 3, E120 – 3.6). Taigi, naudojant 95% klaidingo klasifikavimo tikimybės pasikliautinuosius intervalus, galima atskirti tik ženkliai (kelis kartus) klasifikavimo kokybe besiskiriančius klasifikavimo su mokytoju metodus.

IŠVADOS

1. Išanalizuoti klasifikavimo su mokytoju metodų (tiesinės, kvadratinės, branduolinės, artimiausių kaimynų diskriminantinių analizių, logistinės regresinės analizės, neuroninių tinklų bei klasifikavimo medžių) privalumai bei trūkumai. Nustatyta, kad nėra universalaus metodo įvertinančio klasifikavimo su mokytoju metodų klasifikavimo kokybę. Tam naudojami įvairūs klaidingo klasifikavimo tikimybės taškiniai įverčiai, kurių kiekvienas turi savų privalumų ir trūkumų. Vertinant klasifikavimo kokybę nenaudojami intervaliniai įverčiai. Apibendrinus analizės rezultatus pasiūlyta - parenkant tinkamiausią klasifikavimo su mokytoju metodą naudoti kelis klaidingo klasifikavimo tikimybės taškinius bei intervalinius įverčius.

2. Atlikus SAS sistemos klasifikavimo su mokytoju procedūrų ir makrokomandų analizę, nustatyta, kad lyginant klasifikavimo su mokytoju metodų taikymo rezultatus reikia naudoti skirtingas procedūras bei makrokomandas, kurios pateikia rezultatus įvairiais formatais bei skirtingose duomenų žingsnio programos vietose. Be to keliose SAS procedūrose realizuoti tik du arba visai nerealizuoti (logistinės regresinės analizės atveju) klaidingo klasifikavimo tikimybės vertinimo metodai. Visose procedūrose nerealizuoti klaidingo klasifikavimo tikimybės intervalinių įverčių radimo metodai. Taigi, klasifikavimo rezultatų lyginimas naudojant SAS procedūras ir makrokomandas yra nepatogus, reikalauja didelių darbo sąnaudų, gerų SAS programavimo kalbos žinių ir įgūdžių.

3. Pasiūlyta metodika ir SAS sistemos galimybės išplėstos programiniu klasifikavimo su mokytoju metodų taikymo rezultatų lyginimo įrankiu neturinčiu aukščiau paminėtų trūkumų. Klasifikavimo metodų su mokytoju taikymo rezultatų lyginimui naudojami:

- klaidingo klasifikavimo tikimybės taškiniai įverčiai, gauti savos imties, kryžminio patikrinimo, įkelčių bei Monte Karlo kryžminio patikrinimo metodais;
- klaidingo klasifikavimo tikimybės pasikliautinieji intervalai;
- klaidingo klasifikavimo tikimybės įverčių imčių medianos ir medianų pasikliautinieji intervalai.

Naudojantis sukurtu programiniu įrankiu galima tiksliau ir patogiau lyginti klasifikavimo su mokytoju metodų taikymo skirtingiems duomenimis rezultatus. Šis darbas nereikalauja gerų SAS programavimo kalbos žinių ir didelių darbo sąnaudų. Sukurtas įrankis suteikia vartotojui galimybę parinkti vieną iš 5 klasifikavimo su mokytoju metodų (tiesinės, kvadratinės, branduolinės, artimiausių kaimynų diskriminantinės analizės ir logistinės regresinės analizės), geriausiai tinkantį konkretiems duomenims.

4. Atliktas sukurto įrankio testavimas su skirtingais duomenimis (įvairūs imties dydžiai, skirtingas klasių atskiriamumas, tenkinamos/netenkinamos klasifikavimo metodų taikymo prielaidos ir t.t.) parodė, kad jis sprendžia darbe suformuluotas užduotis. Kiekvienu atveju buvo parinktas tinkamiausias klasifikavimo metodas ir įvertinta jo klasifikavimo kokybė.

5. Sukurtą programinį įrankį galima lengvai papildyti naujais klasifikavimo su mokytoju metodais, (pvz., klasifikavimo medžiais ar neuroniniais tinklais, kurie realizuoti SAS/Enterprise Minner modulyje) ir naujais klaidingo klasifikavimo tikimybės vertinimo metodais.

LITERATŪRA

1. AnswerTree 2.0 Users Guide. SPSS Inc. JAV. 1998, p. 180-197.
2. Bergerud, W. A. Introduction to logistic regression models. Biometrics information Handbook, 1996, No. 7. [Žiūrėta 2006-03-06]. Prieiga per internetą: <http://www.for.gov.bc.ca>.
3. Boik, R.J. Corrections to course Notes: statistics 537 classical multivariate analysis. Spring 2006. [Žiūrėta 2006-03-06]. Prieiga per internetą: <www.math.montana.edu/~rjboik/classes/537/corrections_537.pdf>.
4. Braga-Neto, U. M.; Dougherty, E. R. Is cross-validation valid for small-sample microarray classification? Bioinformatics, Vol. 20, No. 3, 2004, p. 374–380.
5. Braga-Neto, U. M.; Hashimoto, R.; Dougherty, E. R.; Nguyen, D. V.; Carroll, R. J. Is cross-validation better than resubstitution for ranking genes? Bioinformatics, Vol. 20, No. 2, 2004, p. 253–258.
6. Cover, T. M.; Hart, P. E. Nearest neighbor pattern classification. IEEE transaction on information theory, Vol. 13, No. 1, p. 21-27. [Žiūrėta 2004-10-06]. Prieiga per internetą: <http://yreka.stanford.edu/~cover/papers/transIT/0021cove.pdf>.
7. Čekanavičius, V.; Murauskas, G. Statistika ir jos taikymai, II dalis. Vilnius, 2002, 270 p.
8. Feldman, D.; Gross, S. Mortage Default: Classificaton trees analysis. 2003. [Žiūrėta 2004-10-08]. Prieiga per internetą: <http://ideas.repec.org/a/kap/jrefec/v30y2005i4p369-396.html>.
9. Glynn, P.; Sohoni, D.; Leith, L. Multinomial logistic regression using SAS and PROC CATMOD. [Žiūrėta 2004-11-12]. Prieiga per internetą: <http://support.sas.com>.
10. Gutierrez-Pena, E. Bayesian classification methods. Psychology Science, Vol. 46, No 1, 2004, p. 52-64.
11. Hardle, W. Applied nonparametric rgression. Berlin. 1989. [Žiūrėta 2004-10-05]. Prieiga per internetą: <www.quantlet.com/mdstat/scripts/anr/pdf/anrpdf.pdf>.
12. Harris, D. V.; Pan G. Mineral Favorability Mapping: A Comparison of Artificial Neural Networks, Logistic Regression, and Discriminant Analysis. Natural Resources Research, Vol. 8, No. 2, 1999, p. 93-109.
13. Haukoos, J.S.; Lewis, R.J. Advaned statistics: Bootstrapping confidence intervals for statistics with „Difficult“ distributions. Academic emergency medicine, 2005, Vol. 12, No. 4, p. 360-365.
14. Hubert, M.; Driessen, K.V. Fast and robust discriminant analysis. 2002 [Žiūrėta 2005-01-10]. Prieiga per internetą: <www.compstat2004.cuni.cz>.
15. Insightful software and servises. S-Plus. [Interaktyvus]. [Žiūrėta 2006-03-06]. Prieiga per internetą: <http://www.insightful.com/products/splus>.

16. Interpreting Diagnostic Tests. University of Nebraska Medical Center. [Interaktyvus]. [Žiūrėta 2006-02-06]. Prieiga per internetą: <http://gim.unmc.edu/dxtests/Default.htm>.
17. Liu, H.; Wu, T. Estimating the area under Receiver Operating Characteristic (ROC) curve for repeated measures design. [Žiūrėta 2006-03-06]. Prieiga per internetą: <http://www.jstatsoft.org/v08/i12/roc.pdf>.
18. Michie, D.; Spiegelhalter, D.J.; Taylor, C.C. Machine Learning, Neural and Statistical Classification., 1994.
19. Molinaro, A. M.; Simon, R.; Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. Bioinformatics, Vol. 21, No. 15, 2005, p. 3301–3307.
20. Nicolich, M.; Jorgensen, G. Graphical presentation of a nonparametric regression with bootstrapped confidence intervals. Prieiga per internetą: <http://www2.sas.com/proceedings/sugi23/Stats/p248.pdf>.
21. Nothnagel, M. Klassifikationsverfahren der Diskriminanzanalyse, Eine vergleichende und integrierende Übersicht. Diplomarbeit an der Humbolt-Universität zu Berlin Mathematisch-Naturwissenschaftliche Fakultät II Institut für Mathematik, 1999.
22. Pickard, J. M.Using reseiver operating characteristic (ROC) curves to evaluate digital mammography. University of Texas Health Science Center, San Antonio, Texas. [Žiūrėta 2006-02-06]. Prieiga per internetą: <http://ric.uthscsa.edu/personalpages/lancaste/DI2_Projects_2004/JP_Project.pdf>.
23. Pohar, M.; Blas, M.; Turk, S; Comparison of logistic regression and linear discriminant analysis: a simulation study. Metodološki zvezki, 2004, Vol. 1, No. 1, p. 143-161.
24. Quantitative Research in Public Administration PA 765. Logistic Regression. [Interaktyvus]. [Žiūrėta 2005-07-06]. Prieiga per internetą: <http://www2.chass.ncsu.edu/garson/pa765/logistic.htm#nonpar>.
25. Raudys, Š. Statistical and Neural Classifiers: An integrated approach to design. London, 2001.
26. SAS global forum. Sugi 27. Lajiness M. S. A practical introduction to the power of Enterprise Miner. [Interaktyvus]. [Žiūrėta 2006-02-06]. Prieiga per internetą: <http://support.sas.com/events/sasglobalforum>.
27. SAS global forum. Sugi 31, 2006. Gönen, M. Receiver Operating Characteristic (ROC) Curves. Memorial Sloan-Kettering Cancer Center. [Interaktyvus]. [Žiūrėta 2006-02-06]. Prieiga per internetą: <http://support.sas.com/events/sasglobalforum>
28. SAS OnlineDoc. SAS Institute Inc., Cary, NC, 2005.
29. SAS support. Samples SAS/STAT. [Interaktyvus]. [Žiūrėta 2006-03-06]. Prieiga per internetą: <http://support.sas.com/ctx/samples/index.jsp?product=stat&sxf=s&c1=stat&mdType=product&sort=&st=30>

30. Soukup, M.; Cho, H. J.; Lee, J. K. Robust classification modeling on microarray data using misclassification penalized posterior. Bioinformatics, Vol. 21, No. 1, 2005, p. 423–430.
31. SPSS. [Interaktyvus]. [Žiūrėta 2006-03-06]. Prieiga per internetą: <http://www.spss.com>.
32. StatSoft [Interaktyvus]. Electronic Textbook. [Žiūrėta 2004-10-06]. Prieiga per internetą: <http://www.statsoft.com/textbook/stathome.html>.
33. Sueyoshi, T. DEA-Discriminant Analysis: Methodological comparison among eight discriminant analysis approaches. European Journal of Operational Research 169, 2006, p. 247–272.
34. Svantesson, T.;Vallace, J.W. Tests for assessing multivariate normality and the covariance structure of mimo data. Department of Electrical and Computer Engineering Brigham Young University, Provo. [Žiūrėta 2006-03-06]. Prieiga per internetą: <www.ee.byu.edu/wireless/pubs/wallace/icassp_03.pdf>.
35. UCI Machine Learning. [Interaktyvus]. [Žiūrėta 2006-05-03]. Prieiga per internetą: <http://www.ics.uci.edu/~mlearn/MLSummary.html>.
36. Ward Systems Group. Artificial Inteligence Software for Science and Business. [Interaktyvus]. [Žiūrėta 2006-04-06]. Prieiga per internetą: <http://www.wardsystems.com>.
37. Warren, F.K.Graphical Scatter Plots of Labeled Points. [Žiūrėta 2006-04-06]. Prieiga per internetą: <http://support.sas.com/techsup/technote/ts722k.pdf>.
38. Wu, B.; Abbott, T.; Fishman, D.; McMurray, W.; Mor, G.; Stone, K.; Ward, D.; Zhao, K. W. H. Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. Bioinformatics, Vol. 19, No. 13, 2003, p. 1636–1643.
39. Zhang, H. Recursive Partitioning and Tree-based Methods. [Žiūrėta 2004-10-06]. Prieiga per internetą: <http://peace.med.yale.edu>.

1 priedas. Imčių generavimo ir glodinimo parametro parinkimo makrokomandos

Makro komanda ***Gauso_class_3*** skirta trijų dvimačių Gauso skirstinių mišinio generavimui:

$$1: n11 \cdot N\left(\left(m\_x1\_clas1; m\_x2\_clas1\right), \begin{pmatrix} std\_x1\_clas1 & 0 \\ 0 & std\_x1\_clas1 \end{pmatrix} \right) +$$

$$+ n12 \cdot N\left(\left(m\_x1\_clas1\_out; m\_x2\_clas1\_out\right), \begin{pmatrix} std\_x1\_clas1 & 0 \\ 0 & std\_x1\_clas1 \end{pmatrix} \right),$$

$$2: n21 \cdot N\left(\left(m\_x1\_clas2; m\_x2\_clas2\right), \begin{pmatrix} std\_x1\_clas2 & 0 \\ 0 & std\_x1\_clas2 \end{pmatrix} \right) +$$

$$+ n22 \cdot N\left(\left(m\_x1\_clas2\_out; m\_x2\_clas2\_out\right), \begin{pmatrix} std\_x1\_clas2 & 0 \\ 0 & std\_x1\_clas2 \end{pmatrix} \right),$$

$$3: n31 \cdot N\left(\left(m\_x1\_clas3; m\_x2\_clas3\right), \begin{pmatrix} std\_x1\_clas3 & 0 \\ 0 & std\_x1\_clas3 \end{pmatrix} \right) +$$

$$+ n32 \cdot N\left(\left(m\_x1\_clas3\_out; m\_x2\_clas3\_out\right), \begin{pmatrix} std\_x1\_clas3 & 0 \\ 0 & std\_x1\_clas3 \end{pmatrix} \right),$$

čia pirmas skaičius žymi klasę, kurios objektai generuojami; antras skaičius (pirmas po dvitaškio) žymi generuojamų objektų skaičių; po jo nurodytas diskriminavimo kintamųjų skirstinys; po pliuso ženklo vėl nurodomas objektų skaičius bei diskriminavimo kintamųjų skirstinys. Naudojant šią makrokomandą sugeneruotos darbe naudojamos A, B ir C tipo imtys.

Makro komandos kreipinys:

```
%Gauso_class_3(

d_out=duomenys,    m_x1_clas1=0,           m_x1_clas2=0,      m_x1_clas3=0,
x_out=x,           m_x2_clas1=0,           m_x2_clas2=0,      m_x2_clas3=0,
y_out=y,           std_x1_clas1=1,         std_x1_clas2=1,    std_x1_clas3=1,
                   std_x2_clas1=1,         std_x2_clas2=1,    std_x2_clas3=1,
n11=10,
n21=10,            m_x1_clas1_out=0,       m_x1_clas2_out=0   m_x1_clas3_out=0,
n31=10,            m_x2_clas1_out=0,       m_x2_clas2_out=0,  m_x2_clas3_out=0,
n12=0,
n22=0,
n32=0 )
```

Naudojami parametrai:

- *D_out* – duomenų žingsnio programa, kurioje saugomos sugeneruotų atsitiktinių dydžių reikšmės;
- *X_out* – diskriminavimo kintamųjų vardo šaknis. Nurodžius *x_out=x*, bus generuojami tokie diskriminavimo kintamieji: x1, x2.
- *Y_out* – priklausomo kintamojo vardas. Šio kintamojo reikšmės {1, 2, 3}.

Kiti parametrai yra atitinkami mišinį sudarančių skirstinių parametrai.

Makrokomanda ***ZG_3*** skirta trijų dvimačių skirstinių mišinio generavimui:

$1 : n_1 \cdot ZG(0,\ 0,\ m\_z\_clas1,\ 1,\ 1,\ 1),$

$2 : n_2 \cdot ZG(0,\ 0,\ m\_z\_clas2,\ 1,\ 1,\ 1),$

$3 : n_3 \cdot N((0;0), I),$

ZG – žiedinis Gauso skirstinys. Atsitiktinį dydį $Y = (X_1, X_2)$ vadiname Žiediniu dvimačiu Gauso dydžiu, jeigu $X_1 = \frac{U \cdot Z}{\sqrt{U^2 + V^2}}$, $X_2 = \frac{V \cdot Z}{\sqrt{U^2 + V^2}}$. Čia U, V, Z yra vienmačiai atsitiktiniai dydžiai, $U \sim N(\mu_u, \sigma_u^2)$, $V \sim N(\mu_v, \sigma_v^2)$, $Z \sim N(\mu_z, \sigma_z^2)$. Žymime $Y \sim ZG(\mu_u, \mu_v, \mu_z, \sigma_u^2, \sigma_v^2, \sigma_z^2)$. Kiti žymėjimai analogiški naudojamiems makrokomandos *Gauso_class3* atveju. Naudojant makrokomandą *ZG*_3 sugeneruotos darbe naudojamos D ir E tipo imtys.

Makro komandos kreipinys:

```
%ZG_3   (   d_out=duomenys,   m_z_clas1=10,          n1=10,
            x_out=x,          m_z_clas2=5,           n2=10,
            y_out=y,                                 n3=10 )
```

Naudojami parametrai:

- *D_out* – duomenų žingsnio programa, kurioje saugomos sugeneruotų atsitiktinių dydžių reikšmės;
- *X_out* – diskriminavimo kintamųjų vardo šaknis. Nurodžius *x_out=x*, bus generuojami tokie diskriminavimo kintamieji: x1, x2.
- *Y_out* – priklausomo kintamojo vardas. Šio kintamojo reikšmės {1, 2, 3}.

Kiti parametrai yra atitinkami mišinį sudarančių skirstinių parametrai.

Makrokomanda ***Optimal_r*** naudojama branduolio (su penkiom skirtingomis branduolio funkcijomis) ir artimiausių kaimynų metodų glodinimo parametrų parinkimui. Nurodžius pradinę, galutinę glodinimo parametro reikšmes bei žingsnį, kuriuo keičiamas glodinimo parametras, su kiekviena glodinimo parametro reikšme pasirinktais metodais atliekamas klasifikavimas ir kryžminio patikrinimo metodu įvertinama klaidingo klasifikavimo tikimybė. Vartotojui išvedami glodinimo parametrai bei su jais gauti klaidingo klasifikavimo tikimybės taškiniai kryžminio patikrinimo įverčiai.

Makro komandos kreipinys:

```
%optimal_r (d_in=,      start_r=0,  stop_r=0,   step_r=0,   d_out=parameter,
            KDA_biw=0,  KDA_epa=0,  KDA_norm=0, KDA_tri=0, KDA_uni=0, NNDA=0).
```

Naudojami parametrai:

- *D_in* – duomenų žingsnio programa, kurioje saugomi pradiniai duomenys (mokomoji imtis).
- *Start_r* – pradinė glodinimo parametro reikšmė.
- *Stop_r* – galutinė glodinimo parametro reikšmė.
- *Step_r* – žingsnis, kuriuo keičiama glodinimo parametro reikšmė.
- *D_out* – duomenų žingsnio programa, kurioje saugomi rezultatai, t.y. klasifikavimo su mokytoju metodo pavadinimai, glodinimo parametro reikšmės ir klaidingo klasifikavimo tikimybės taškinis kryžminio patikrinimo įvertis.
- *KDA_biw* – klasifikavimui naudojamas branduolinės diskriminantinės analizės su dvisvoriu branduoliu metodas.
- *KDA_epa* – klasifikavimui naudojamas branduolinės diskriminantinės analizės su Epanechnikovo branduoliu metodas.
- *KDA_norm* – klasifikavimui naudojamas branduolinės diskriminantinės analizės su normaliuoju branduoliu metodas.
- *KDA_tri* – klasifikavimui naudojamas branduolinės diskriminantinės analizės su trisvoriu branduoliu metodas.
- *KDA_uni* – klasifikavimui naudojamas branduolinės diskriminantinės analizės su tolygiuoju branduoliu metodas.
- *NNDA* – klasifikavimui naudojamas artimiausių kaimynų diskriminantinė analizė metodas.

2 priedas. Eksperimentų rezultatų lentelės

19 lentelė

Testinės imties klasifikavimo rezultatai

BDA metodu (imtis D120)

Number of Observations and Percent Classified into y					
	class_1	**class_2**	**class_3**	**Other**	**Total**
Total	5368 36.66	3940 26.91	4266 29.14	1067 7.29	14641 100.00
Priors	0.33333	0.33333	0.33333		

20 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A30)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0000	0.0333	0.0774	0.0000	0.2667	0.0508
2	BDA (Epanechnikovo f-ja)	0.0667	0.0667	0.1497	0.0000	0.3889	0.1204
3	BDA (dvisvore f-ja)	0.0000	0.0667	0.1480	0.0000	0.3889	0.1108
4	BDA (normalioji f-ja)	0.0000	0.0333	0.0588	0.0000	0.2222	0.0653
5	BDA (tolygioji f-ja)	0.1000	0.1000	0.1879	0.0000	0.4222	0.1630
6	BDA (trisvore f-ja)	0.0000	0.0667	0.1390	0.0000	0.3889	0.1078
7	KDA	0.0667	0.1000	0.1544	0.0000	0.3889	0.1118
8	LR	0.0000	0.1333	0.1768	0.0714	0.4444	0.1583
9	TDA	0.0667	0.0667	0.0878	0.0000	0.2500	0.0874

21 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A30)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.0667	0.0556	0.0833	0.0000	0.0000	0.0000
2	BDA (Epanechnikovo f-ja)	0.1333	0.0833	0.1778	0.1111	0.0000	0.1250
3	BDA (dvisvore f-ja)	0.1500	0.1333	0.1667	0.1111	0.0833	0.1250
4	BDA (normalioji f-ja)	0.0667	0.0000	0.0667	0.0000	0.0000	0.1111
5	BDA (tolygioji f-ja)	0.1222	0.0833	0.1667	0.1111	0.0000	0.1111
6	BDA (trisvore f-ja)	0.1861	0.1667	0.2222	0.1667	0.1250	0.1667
7	KDA	0.1667	0.1333	0.1667	0.1111	0.0000	0.1250

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
8	LR	0.1667	0.1333	0.1818	0.1667	0.1667	0.1667
9	TDA	0.0833	0.0667	0.1111	0.0000	0.0000	0.1111

22 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A120)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0917	0.1500	0.1785	0.0979	0.2719	0.1505
2	BDA (Epanechnikovo f-ja)	0.1167	0.1167	0.1240	0.0541	0.2164	0.1039
3	BDA (dvisvore f-ja)	0.1250	0.1250	0.1392	0.0513	0.2248	0.1169
4	BDA (normalioji f-ja)	0.1167	0.1250	0.1336	0.0641	0.2216	0.1134
5	BDA (tolygioji f-ja)	0.1167	0.1250	0.1385	0.0641	0.2471	0.1212
6	BDA (trisvore f-ja)	0.1167	0.1250	0.1417	0.0667	0.2216	0.1232
7	KDA	0.1167	0.1333	0.1257	0.0452	0.2286	0.1125
8	LR	0.1167	0.1333	0.1304	0.0652	0.2174	0.1121
9	TDA	0.1167	0.1167	0.1185	0.0452	0.1915	0.1071

23 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A120)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.1775	0.1667	0.1905	0.1619	0.1310	0.1739
2	BDA (Epanechnikovo f-ja)	0.1416	0.1357	0.1526	0.1273	0.1157	0.1558
3	BDA (dvisvore f-ja)	0.1347	0.1228	0.1481	0.1263	0.1111	0.1306
4	BDA (normalioji f-ja)	0.1355	0.1255	0.1443	0.1217	0.1082	0.1296
5	BDA (tolygioji f-ja)	0.1430	0.1357	0.1578	0.1279	0.1157	0.1576
6	BDA (trisvore f-ja)	0.1361	0.1212	0.1481	0.1172	0.1000	0.1286
7	KDA	0.1315	0.1212	0.1389	0.1259	0.1037	0.1286
8	LR	0.1250	0.1163	0.1429	0.1250	0.1250	0.1250
9	TDA	0.1277	0.1148	0.1381	0.1172	0.1032	0.1286

24 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A300)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.1267	0.1400	0.16618	0.10989	0.22088	0.14057
2	BDA	0.1267	0.1300	0.14461	0.10077	0.20486	0.13124
3	KDA	0.1300	0.1367	0.14232	0.09800	0.20967	0.13226
4	LR	0.1267	0.1367	0.13978	0.08929	0.20690	0.12967
5	TDA	0.1267	0.1300	0.13926	0.09487	0.19475	0.12650

25 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis A300)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.1692	0.1628	0.1721	0.1358	0.1258	0.1492
2	BDA	0.1430	0.1356	0.1522	0.1270	0.1190	0.1369
3	KDA	0.1400	0.1347	0.1465	0.1341	0.1222	0.1413
4	LR	0.1366	0.1313	0.1442	0.1333	0.1167	0.1333
5	TDA	0.1382	0.1310	0.1443	0.1226	0.1174	0.1341

26 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B30)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.2333	0.3333	0.43069	0.16667	0.66667	0.40872
2	BDA	0.3667	0.3667	0.44183	0.16667	0.66667	0.40728
3	KDA	0.1667	0.3333	0.40329	0.11111	0.77778	0.33283
4	LR	0.3667	0.6667	0.57599	0.36364	0.85714	0.59667
5	TDA	0.3667	0.5667	0.48666	0.25000	0.66667	0.53292

27 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B30)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.4167	0.3889	0.4444	0.4306	0.3333	0.5000
2	BDA	0.4444	0.4000	0.5000	0.4444	0.3750	0.5000
3	KDA	0.3917	0.3611	0.4444	0.3333	0.2778	0.3333
4	LR	0.5635	0.5385	0.6154	0.6667	0.5000	0.6667
5	TDA	0.5000	0.4722	0.5333	0.5000	0.5000	0.5556

28 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B120)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.2833	0.3083	0.36934	0.26667	0.50794	0.40059
2	BDA	0.1750	0.3500	0.40272	0.29444	0.50476	0.39727
3	KDA	0.3250	0.4000	0.36419	0.24777	0.50553	0.36565
4	LR	0.4417	0.4750	0.51550	0.36735	0.74419	0.49583
5	TDA	0.4250	0.4417	0.44278	0.33810	0.54444	0.45444

29 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B120)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.3625	0.3524	0.3796	0.4067	0.3750	0.4352
2	BDA	0.4068	0.3913	0.4167	0.4024	0.3810	0.4222
3	KDA	0.3646	0.3502	0.3798	0.3615	0.3472	0.3795
4	LR	0.5000	0.4706	0.5238	0.5000	0.4583	0.5417
5	TDA	0.4435	0.4340	0.4569	0.4583	0.4333	0.4881

30 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B300)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.2467	0.3533	0.39492	0.32353	0.48312	0.37818
2	BDA	0.2767	0.3933	0.39285	0.33068	0.45430	0.39218
3	KDA	0.3400	0.3600	0.36226	0.30028	0.41481	0.35505
4	LR	0.4100	0.4300	0.42959	0.35398	0.51282	0.42750
5	TDA	0.4133	0.4200	0.41612	0.35219	0.49729	0.41873

31 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis B300)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.3917	0.3844	0.4033	0.3730	0.3644	0.3841
2	BDA	0.3932	0.3819	0.4026	0.3936	0.3833	0.4042
3	KDA	0.3640	0.3563	0.3728	0.3581	0.3454	0.3667
4	LR	0.4313	0.4234	0.4425	0.4250	0.4000	0.4333
5	TDA	0.4145	0.4069	0.4224	0.4195	0.4100	0.4295

32 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis C130)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0917	0.0917	0.14634	0.053419	0.25406	0.12672
2	BDA	0.0833	0.0944	0.12877	0.044160	0.21852	0.10515
3	KDA	0.1306	0.1306	0.15247	0.057143	0.23838	0.13789
4	LR	0.1000	0.1300	0.13369	0.054054	0.25714	0.12650
5	TDA	0.1306	0.1306	0.15065	0.039216	0.31313	0.13017

33 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis C130)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.1389	0.1286	0.1502	0.1071	0.0847	0.1222
2	BDA	0.1213	0.1124	0.1329	0.0847	0.0833	0.1037
3	KDA	0.1378	0.1245	0.1538	0.1111	0.0952	0.1296
4	LR	0.1212	0.1143	0.1351	0.1000	0.1000	0.1500
5	TDA	0.1505	0.1255	0.1624	0.1194	0.0972	0.1500

34 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D30)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.1333	0.1000	0.25939	0.04762	0.55556	0.17056
2	BDA	0.0667	0.1000	0.25578	0.00000	0.50000	0.16431
3	KDA	0.1000	0.1667	0.24940	0.00000	0.55556	0.18681
4	LR	0.3000	0.46667	0.43065	0.16667	0.77778	0.42500
5	TDA	0.2333	0.3667	0.36851	0.08333	0.63889	0.37778

35 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D30)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.2333	0.2000	0.2778	0.1667	0.1111	0.1667
2	BDA	0.2440	0.2167	0.2778	0.1111	0.1111	0.1667
3	KDA	0.2556	0.2000	0.3000	0.1667	0.1111	0.1667
4	LR	0.4444	0.4000	0.4545	0.5000	0.3333	0.5000
5	TDA	0.3631	0.3333	0.3889	0.3542	0.3333	0.4444

36 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D300)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0033	0.0067	0.02020	0.00000	0.04580	0.00912
2	BDA	0.0033	0.0100	0.02963	0.00000	0.06916	0.01430
3	KDA	0.0167	0.0233	0.04177	0.01333	0.09042	0.02680
4	LR	0.6167	0.83333	0.65077	0.45192	0.79339	0.66883
5	TDA	0.6167	0.7100	0.61232	0.53954	0.69606	0.64311

37 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis D300)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.0190	0.0175	0.0248	0.0000	0.0000	0.0133
2	BDA	0.0273	0.0255	0.0317	0.0133	0.0000	0.0159
3	KDA	0.0362	0.0303	0.0447	0.0208	0.0175	0.0278
4	LR	0.6652	0.6455	0.6842	0.6667	0.6667	0.7000
5	TDA	0.6119	0.6037	0.6213	0.6396	0.6295	0.6496

38 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E30)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.3333	0.4667	0.56494	0.20000	0.86667	0.57836
2	BDA	0.3333	0.4000	0.47477	0.22222	0.71429	0.44497
3	KDA	0.2333	0.3333	0.38442	0.11111	0.65000	0.34858
4	LR	0.5000	0.9000	0.73598	0.45455	1.00000	0.74500
5	TDA	0.5000	0.7333	0.65038	0.33333	0.91667	0.65269

39 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E30)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.5833	0.5444	0.6111	0.6667	0.6667	0.7222
2	BDA	0.4556	0.4167	0.5000	0.4444	0.3333	0.5000
3	KDA	0.3889	0.3556	0.4333	0.3333	0.2778	0.3333
4	LR	0.7417	0.7000	0.7778	0.8333	0.8333	0.8333
5	TDA	0.6667	0.6111	0.6944	0.7222	0.6667	0.7778

40 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E120)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.0889	0.2000	0.31318	0.12963	0.49584	0.23413
2	BDA	0.1222	0.1778	0.26708	0.11111	0.39583	0.21073
3	KDA	0.1111	0.2000	0.22586	0.06667	0.41026	0.18485
4	LR	0.5333	0.77778	0.68464	0.45946	0.87879	0.68944
5	TDA	0.5333	0.6556	0.63924	0.46434	0.77910	0.65857

41 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E120)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.3146	0.3007	0.3379	0.2347	0.2222	0.2540
2	BDA	0.2663	0.2606	0.2833	0.2171	0.1905	0.2361
3	KDA	0.2230	0.1998	0.2452	0.1831	0.1667	0.2083
4	LR	0.6821	0.6571	0.7097	0.7222	0.6667	0.7222
5	TDA	0.6350	0.6111	0.6629	0.6538	0.6397	0.6786

42 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E300)

Nr.	Metodas	SI	KP	IK	IK_95a	IK_95v	MKKP
1	AKDA	0.1800	0.2200	0.26678	0.19666	0.34481	0.23428
2	BDA	0.1767	0.2167	0.25294	0.18481	0.31915	0.22435
3	KDA	0.1967	0.2100	0.23153	0.15310	0.33365	0.20865
4	LR	0.6100	0.6633	0.68607	0.54000	0.81651	0.69650
5	TDA	0.6100	0.6267	0.65781	0.58354	0.74643	0.65843

43 lentelė

Klaidingo klasifikavimo tikimybės įverčiai (imtis E300)

Nr.	Metodas	IK_me	IK_me_95a	IK_me_95v	MKKP_me	MKKP_me_95a	MKKP_me_95v
1	AKDA	0.2663	0.2582	0.2781	0.2309	0.2210	0.2489
2	BDA	0.2540	0.2466	0.2619	0.2201	0.2101	0.2456
3	KDA	0.2264	0.2193	0.2372	0.1997	0.1920	0.2167
4	LR	0.6898	0.6667	0.7232	0.7167	0.6833	0.7333
5	TDA	0.6531	0.6461	0.6669	0.6657	0.6493	0.6834

44 lentelė

Normalumo tikrinimas

(imtis A30, 1-klasė)

Normality Test			
Equation	Test Statistic	Value	Prob
x1	Shapiro-Wilk W	0.90	0.2284
x2	Shapiro-Wilk W	0.85	0.0512
System	Mardia Skewness	3.68	0.4507
	Mardia Kurtosis	-1.12	0.2646
	Henze-Zirkler T	1.37	0.1709

45 lentelė

Normalumo tikrinimas

(imtis A30, 2-klasė)

Normality Test			
Equation	Test Statistic	Value	Prob
x1	Shapiro-Wilk W	0.98	0.9557
x2	Shapiro-Wilk W	0.92	0.3593
System	Mardia Skewness	7.54	0.1100
	Mardia Kurtosis	0.13	0.8938
	Henze-Zirkler T	-0.24	0.8108

46 lentelė

Normalumo tikrinimas (imtis A30, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.87	0.1045
x2	Shapiro-Wilk W	0.89	0.1700
System	Mardia Skewness	0.92	0.9210
	Mardia Kurtosis	-1.01	0.3114
	Henze-Zirkler T	0.23	0.8158

47 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (A30)

Chi-Square	DF	Pr > ChiSq
5.650788	6	0.4634

48 lentelė

Normalumo tikrinimas (imtis A120, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.9076
x2	Shapiro-Wilk W	0.99	0.9787
System	Mardia Skewness	0.51	0.9724
	Mardia Kurtosis	0.21	0.8365
	Henze-Zirkler T	0.23	0.8188

49 lentelė

Normalumo tikrinimas (imtis A120, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.8948
x2	Shapiro-Wilk W	0.95	0.1536
System	Mardia Skewness	10.24	0.0366
	Mardia Kurtosis	-0.01	0.9920
	Henze-Zirkler T	0.21	0.8308

50 lentelė

Normalumo tikrinimas (imtis A120, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.97	0.4811
x2	Shapiro-Wilk W	0.97	0.3829
System	Mardia Skewness	2.88	0.5780
	Mardia Kurtosis	0.12	0.9008
	Henze-Zirkler T	-0.47	0.6419

51 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (A120)

Chi-Square	DF	Pr > ChiSq
3.942488	6	0.6845

52 lentelė

Normalumo tikrinimas

(imtis A300, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.7833
x2	Shapiro-Wilk W	0.97	0.2040
System	Mardia Skewness	4.36	0.3597
	Mardia Kurtosis	-0.67	0.4997
	Henze-Zirkler T	0.69	0.4928

53 lentelė

Normalumo tikrinimas

(imtis A300, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.4074
x2	Shapiro-Wilk W	0.98	0.5282
System	Mardia Skewness	4.05	0.3999
	Mardia Kurtosis	-1.10	0.2724
	Henze-Zirkler T	0.35	0.7299

54 lentelė

Normalumo tikrinimas

(imtis A300, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.3172
x2	Shapiro-Wilk W	0.97	0.2474
System	Mardia Skewness	1.43	0.8395
	Mardia Kurtosis	-1.76	0.0790
	Henze-Zirkler T	-0.33	0.7431

55 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (A300)

Chi-Square	DF	Pr > ChiSq
8.669001	6	0.1931

56 lentelė

Normalumo tikrinimas

(imtis B30, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.94	0.5838
x2	Shapiro-Wilk W	0.92	0.3731
System	Mardia Skewness	2.10	0.7168
	Mardia Kurtosis	-0.80	0.4241
	Henze-Zirkler T	-0.63	0.5281

57 lentelė

Normalumo tikrinimas

(imtis B30, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.9759
x2	Shapiro-Wilk W	0.88	0.1240
System	Mardia Skewness	7.73	0.1021
	Mardia Kurtosis	-0.21	0.8375
	Henze-Zirkler T	0.48	0.6290

58 lentelė

Normalumo tikrinimas (imtis B30, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
	Shapiro-Wilk W	0.95	0.7009
	Shapiro-Wilk W	0.95	0.7087
	Mardia Skewness	0.87	0.9291
	Mardia Kurtosis	-1.31	0.1913
	Henze-Zirkler T	0.24	0.8101

59 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (B30)

Chi-Square	DF	Pr > ChiSq
41.271002	6	<.0001

60 lentelė

Normalumo tikrinimas (imtis B120, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.99	0.9761
x2	Shapiro-Wilk W	0.96	0.2404
System	Mardia Skewness	4.48	0.3444
	Mardia Kurtosis	0.60	0.5457
	Henze-Zirkler T	0.73	0.4643

61 lentelė

Normalumo tikrinimas (imtis B120, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.96	0.1656
x2	Shapiro-Wilk W	0.96	0.2205
System	Mardia Skewness	0.49	0.9741
	Mardia Kurtosis	-1.99	0.0467
	Henze-Zirkler T	2.26	0.0238

62 lentelė

Normalumo tikrinimas (imtis B120, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.7313
x2	Shapiro-Wilk W	0.97	0.3649
System	Mardia Skewness	4.04	0.4009
	Mardia Kurtosis	1.70	0.0885
	Henze-Zirkler T	1.69	0.0906

63 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (B120)

Chi-Square	DF	Pr > ChiSq
168.840619	6	<.0001

64 lentelė

Normalumo tikrinimas

(imtis B300, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.6355
x2	Shapiro-Wilk W	0.98	0.3923
System	Mardia Skewness	12.77	0.0125
	Mardia Kurtosis	-0.18	0.8587
	Henze-Zirkler T	0.21	0.8338

65 lentelė

Normalumo tikrinimas

(imtis B300, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.99	0.7982
x2	Shapiro-Wilk W	0.97	0.1488
System	Mardia Skewness	1.78	0.7763
	Mardia Kurtosis	-0.35	0.7275
	Henze-Zirkler T	-1.34	0.1786

66 lentelė

Normalumo tikrinimas

(imtis B300, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.98	0.6347
x2	Shapiro-Wilk W	0.98	0.3535
System	Mardia Skewness	3.53	0.4732
	Mardia Kurtosis	-1.02	0.3056
	Henze-Zirkler T	0.48	0.6292

67 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (B300)

Chi-Square	DF	Pr > ChiSq
407.867594	6	<.0001

68 lentelė

Normalumo tikrinimas

(imtis C130, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.94	0.1180
x2	Shapiro-Wilk W	0.96	0.3172
System	Mardia Skewness	3.82	0.4315
	Mardia Kurtosis	0.07	0.9475
	Henze-Zirkler T	0.67	0.5001

69 lentelė

Normalumo tikrinimas

(imtis C130, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.83	<.0001
x2	Shapiro-Wilk W	0.92	0.0099
System	Mardia Skewness	9.08	0.0591
	Mardia Kurtosis	-1.05	0.2915
	Henze-Zirkler T	3.64	0.0003

70 lentelė

Normalumo tikrinimas (imtis C130, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.95	0.2433
x2	Shapiro-Wilk W	0.98	0.8434
System	Mardia Skewness	0.32	0.9886
	Mardia Kurtosis	-1.02	0.3094
	Henze-Zirkler T	0.04	0.9649

71 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (C130)

Chi-Square	DF	Pr > ChiSq
63.331590	6	<.0001

72 lentelė

Normalumo tikrinimas (imtis D30, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.87	0.0999
x2	Shapiro-Wilk W	0.92	0.3538
System	Mardia Skewness	5.10	0.2776
	Mardia Kurtosis	-0.65	0.5135
	Henze-Zirkler T	1.20	0.2285

73 lentelė

Normalumo tikrinimas (imtis D30, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.90	0.1845
x2	Shapiro-Wilk W	0.96	0.8162
System	Mardia Skewness	3.88	0.4232
	Mardia Kurtosis	-1.15	0.2507
	Henze-Zirkler T	1.09	0.2742

74 lentelė

Normalumo tikrinimas (imtis D30, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.75	0.0045
x2	Shapiro-Wilk W	0.86	0.0751
System	Mardia Skewness	9.92	0.0418
	Mardia Kurtosis	-0.07	0.9472
	Henze-Zirkler T	1.58	0.1146

75 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (D30)

Chi-Square	DF	Pr > ChiSq
31.334905	6	<.0001

76 lentelė

Normalumo tikrinimas

(imtis D300, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.89	<.0001
x2	Shapiro-Wilk W	0.91	<.0001
System	Mardia Skewness	0.76	0.9433
	Mardia Kurtosis	-4.68	<.0001
	Henze-Zirkler T	7.06	<.0001

77 lentelė

Normalumo tikrinimas

(imtis D300, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.91	<.0001
x2	Shapiro-Wilk W	0.93	<.0001
System	Mardia Skewness	0.54	0.9691
	Mardia Kurtosis	-4.44	<.0001
	Henze-Zirkler T	6.51	<.0001

78 lentelė

Normalumo tikrinimas

(imtis D300, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.97	0.1090
x2	Shapiro-Wilk W	0.98	0.5137
System	Mardia Skewness	6.81	0.1465
	Mardia Kurtosis	-0.56	0.5746
	Henze-Zirkler T	-0.31	0.7559

79 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (D300)

Chi-Square	DF	Pr > ChiSq
573.931877	6	<.0001

80 lentelė

Normalumo tikrinimas

(imtis E30, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.95	0.6713
x2	Shapiro-Wilk W	0.88	0.1194
System	Mardia Skewness	1.30	0.8608
	Mardia Kurtosis	-1.39	0.1650
	Henze-Zirkler T	1.27	0.2035

81 lentelė

Normalumo tikrinimas

(imtis E30, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.99	0.9857
x2	Shapiro-Wilk W	0.95	0.6903
System	Mardia Skewness	0.74	0.9460
	Mardia Kurtosis	-0.71	0.4787
	Henze-Zirkler T	-1.39	0.1655

82 lentelė

Normalumo tikrinimas (imtis E30, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.93	0.4250
x2	Shapiro-Wilk W	0.94	0.5483
System	Mardia Skewness	5.64	0.2277
	Mardia Kurtosis	0.02	0.9814
	Henze-Zirkler T	0.32	0.7527

83 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (E30)

Chi-Square	DF	Pr > ChiSq
29.685637	6	<.0001

84 lentelė

Normalumo tikrinimas (imtis E120, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.90	0.0068
x2	Shapiro-Wilk W	0.91	0.0136
System	Mardia Skewness	0.40	0.9827
	Mardia Kurtosis	-2.31	0.0209
	Henze-Zirkler T	2.85	0.0044

85 lentelė

Normalumo tikrinimas (imtis E120, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.94	0.0845
x2	Shapiro-Wilk W	0.97	0.4984
System	Mardia Skewness	1.53	0.8206
	Mardia Kurtosis	-1.67	0.0946
	Henze-Zirkler T	1.29	0.1979

86 lentelė

Normalumo tikrinimas (imtis E120, 3-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.96	0.3059
x2	Shapiro-Wilk W	0.96	0.4211
System	Mardia Skewness	0.81	0.9375
	Mardia Kurtosis	-1.34	0.1789
	Henze-Zirkler T	-0.61	0.5436

87 lentelė

Kovariacijų matricų homogeniškumo tikrinimas (E120)

Chi-Square	DF	Pr > ChiSq
105.024395	6	<.0001

88 lentelė

Normalumo tikrinimas

(imtis E300, 1-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.92	<.0001
x2	Shapiro-Wilk W	0.90	<.0001
System	Mardia Skewness	0.18	0.9961
	Mardia Kurtosis	-4.43	<.0001
	Henze-Zirkler T	6.39	<.0001

89 lentelė

Normalumo tikrinimas

(imtis E300, 2-klasė)

Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.97	0.1040
x2	Shapiro-Wilk W	0.98	0.6129
System	Mardia Skewness	1.69	0.7925
	Mardia Kurtosis	-2.50	0.0126
	Henze-Zirkler T	3.03	0.0024

90 lentelė

Normalumo tikrinimas

(imtis E300, 3-klasė)

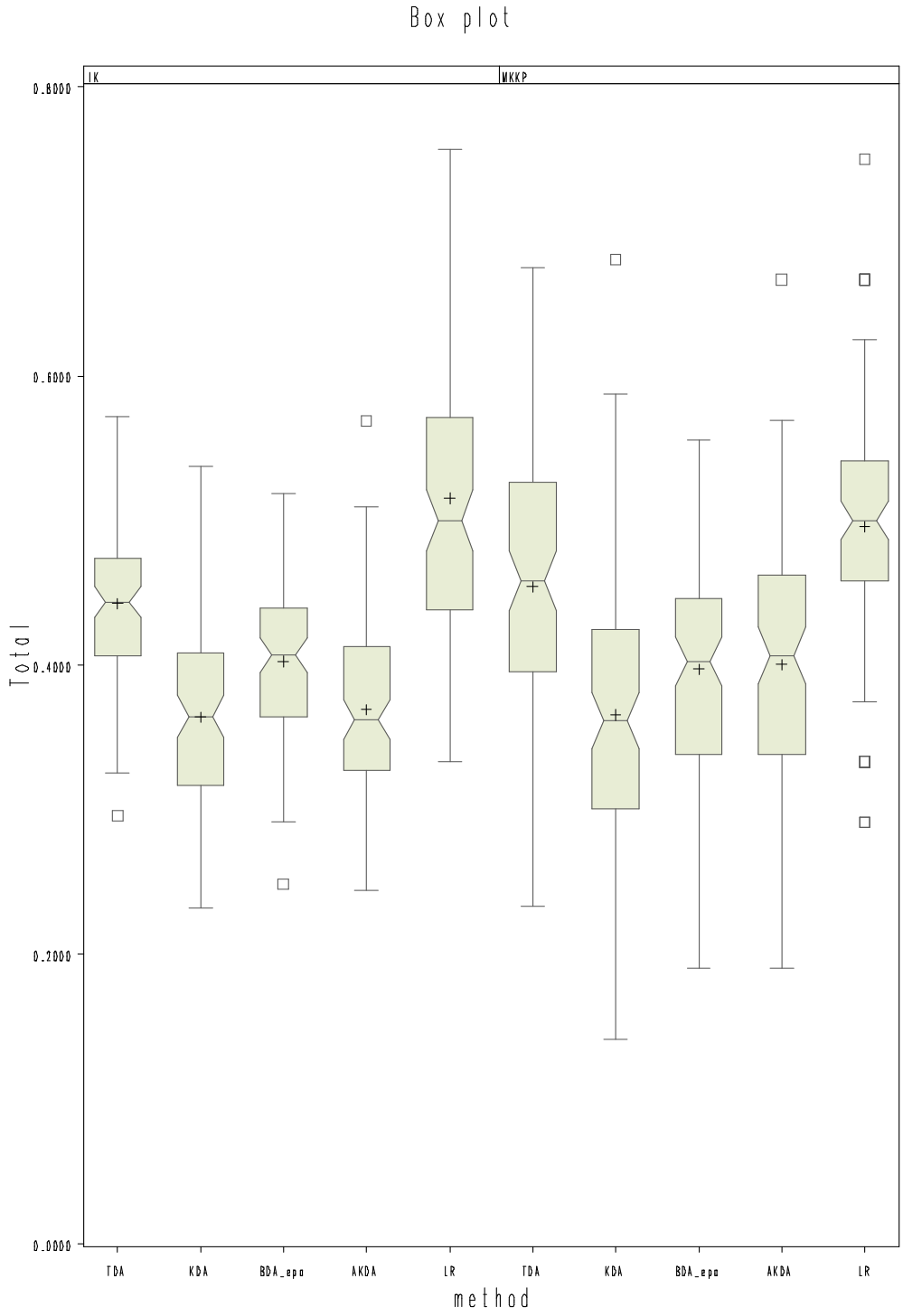
Normality Test			
Equation	**Test Statistic**	**Value**	**Prob**
x1	Shapiro-Wilk W	0.96	0.0160
x2	Shapiro-Wilk W	0.99	0.9301
System	Mardia Skewness	1.84	0.7657
	Mardia Kurtosis	-0.44	0.6624
	Henze-Zirkler T	-0.60	0.5495

91 lentelė

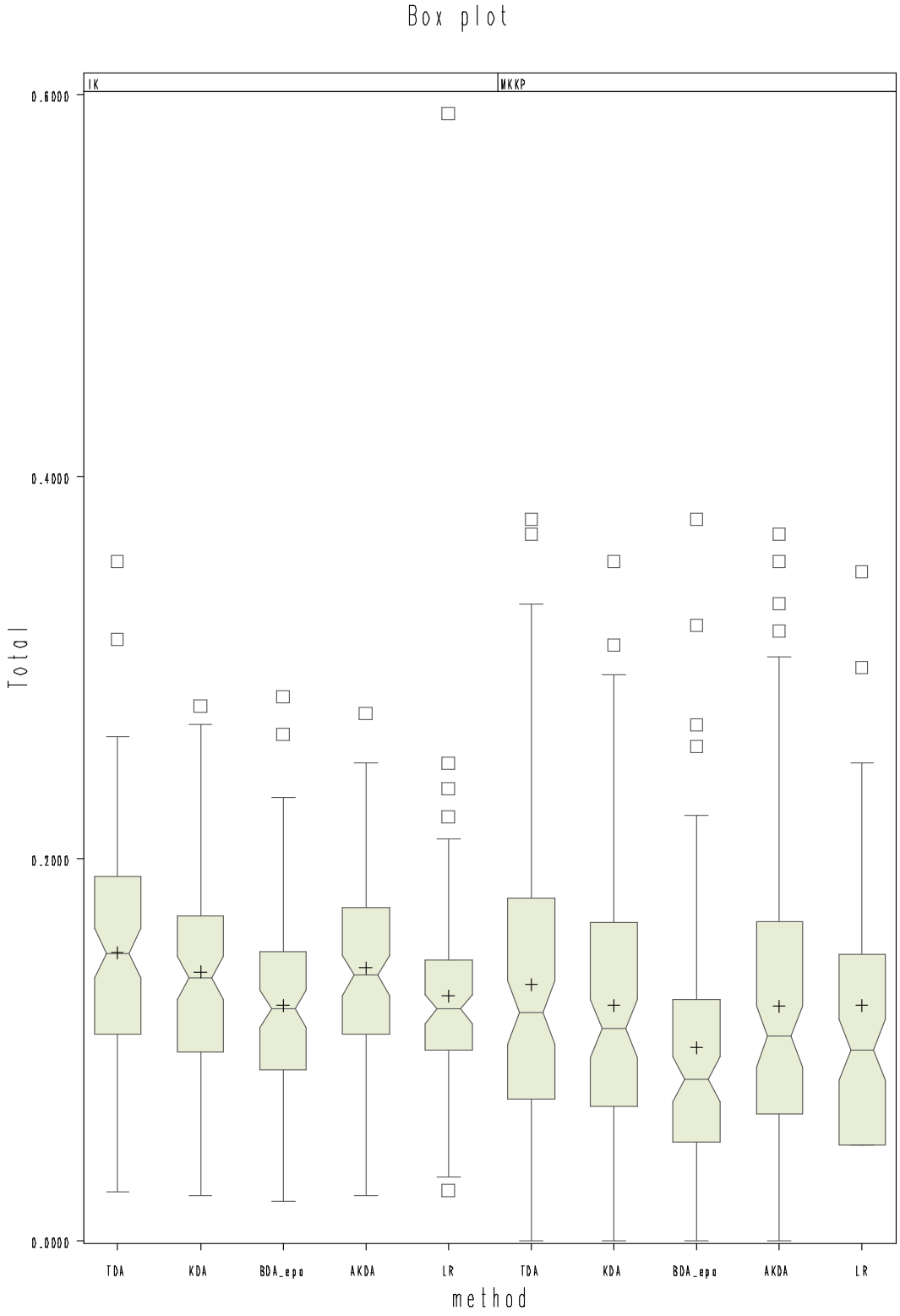
Kovariacijų matricų homogeniškumo tikrinimas (E300)

Chi-Square	DF	Pr > ChiSq
266.472851	6	<.0001

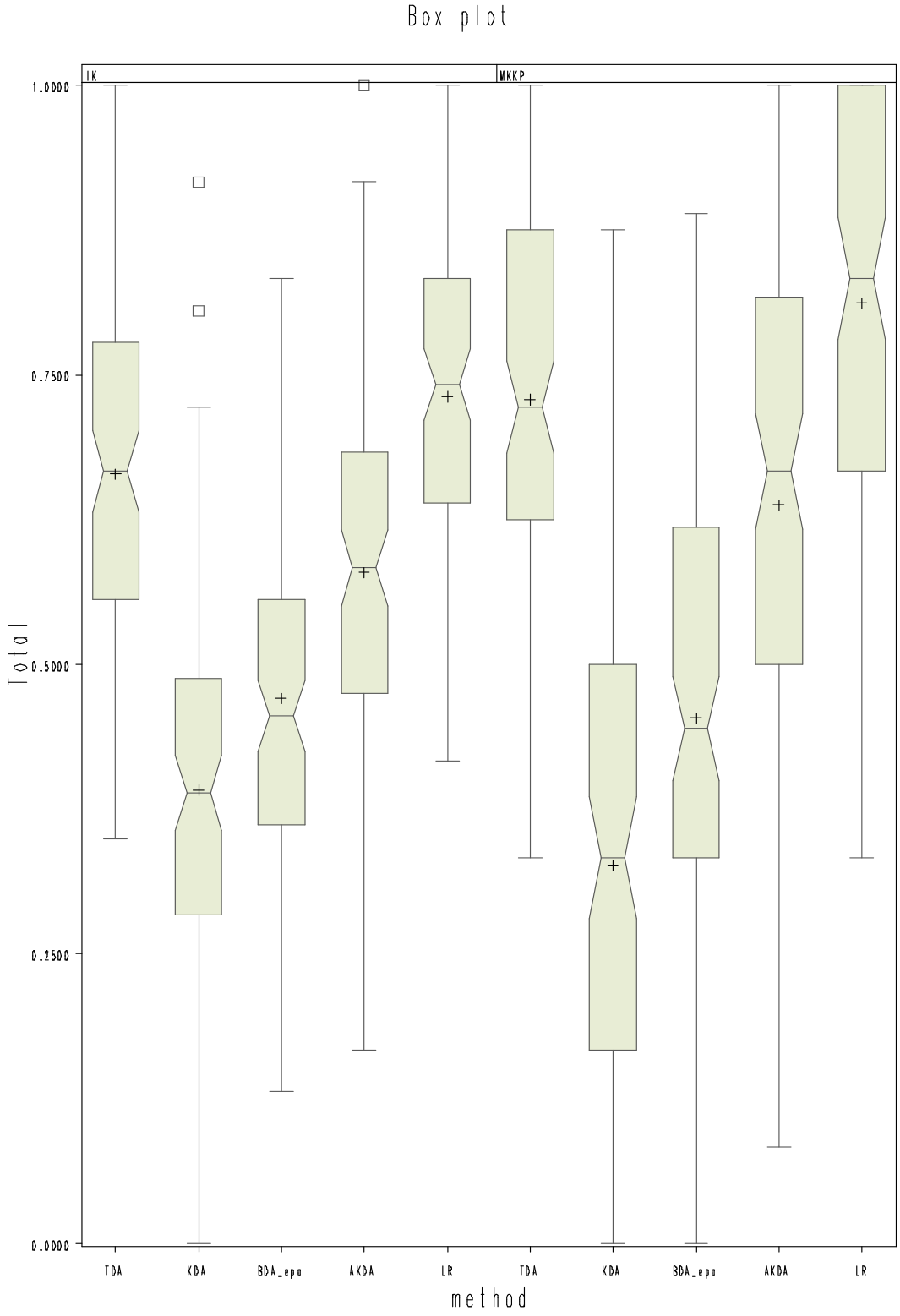
3 priedas. Eksperimentų rezultatų paveikslai



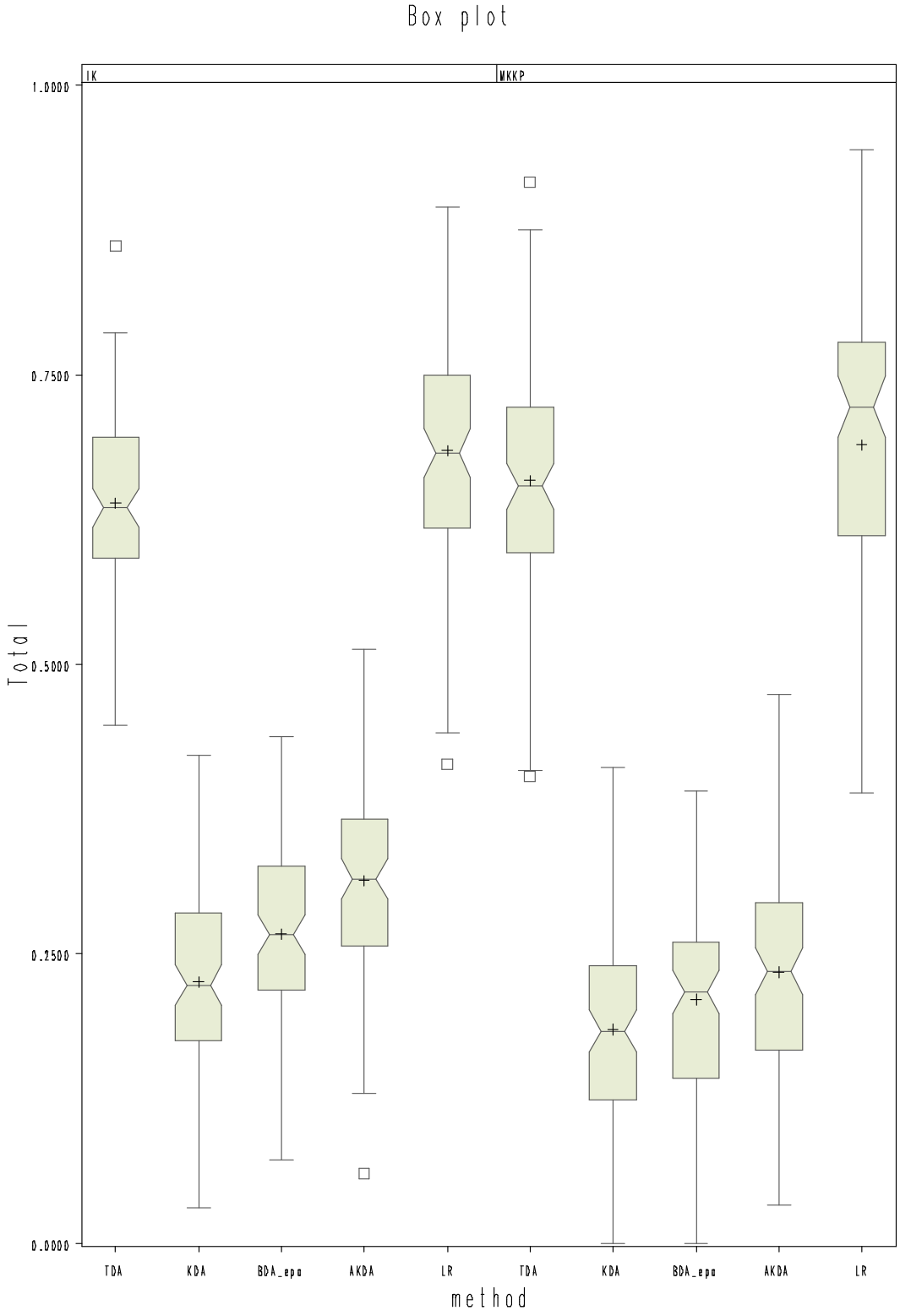
18 pav.Stačiakampės diagramos (imtis B120)



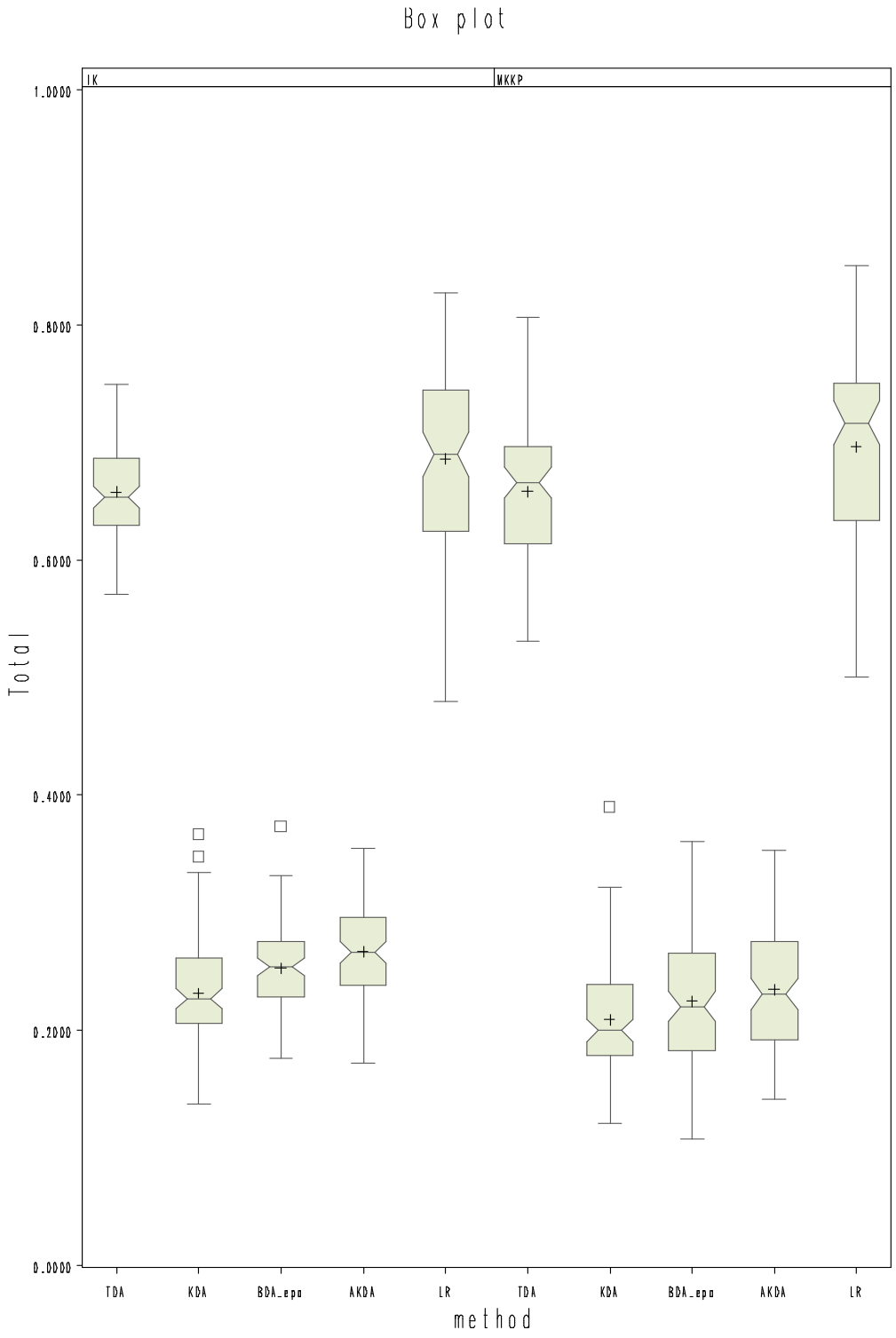
19 pav.Stačiakampės diagramos (imtis C130)



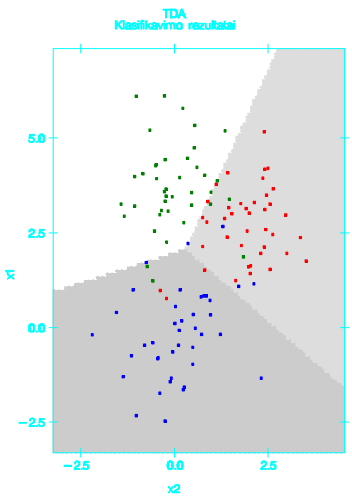
20 pav.Stačiakampės diagramos (imtis E30)



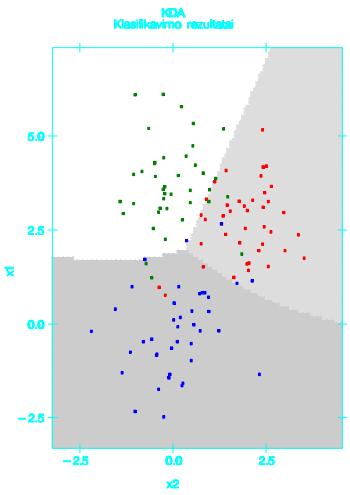
21 pav.Stačiakampės diagramos (imtis E120)



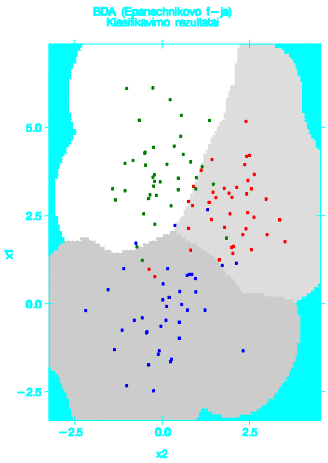
22 pav.Stačiakampės diagramos (imtis E300)



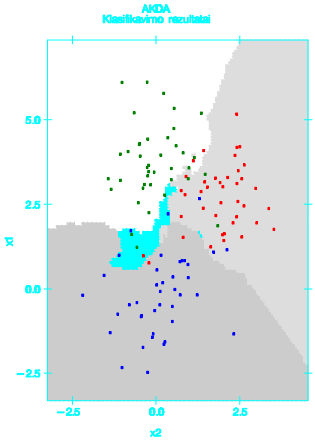
23 pav.Klasifikavimo TDA metodu rezultatai (imtis A120)



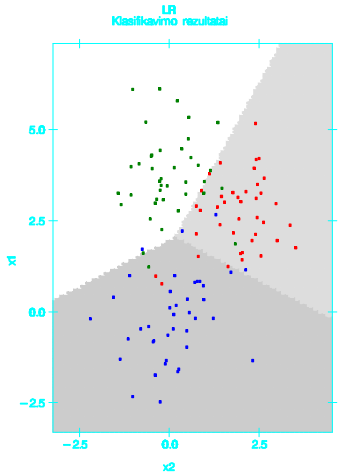
24 pav.Klasifikavimo KDA metodu rezultatai (imtis A120)



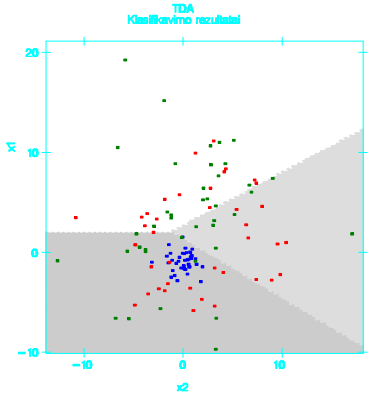
25 pav.Klasifikavimo BDA metodu rezultatai (imtis A120)



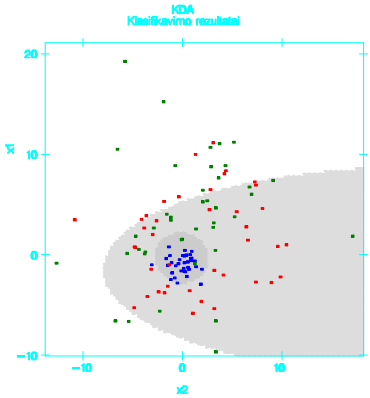
26 pav.Klasifikavimo AKDA metodu rezultatai (imtis A120)



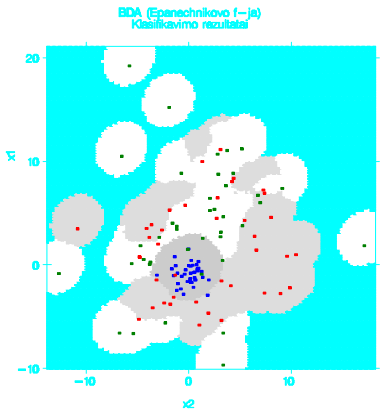
27 pav.Klasifikavimo LR metodu rezultatai (imtis A120)



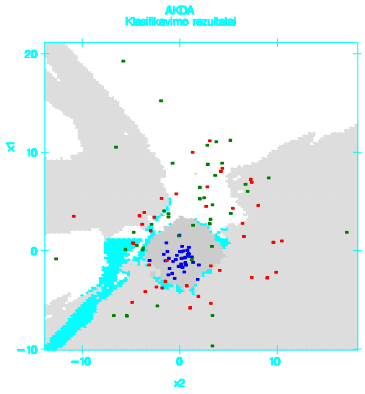
28 pav.Klasifikavimo TDA metodu rezultatai (imtis B120)



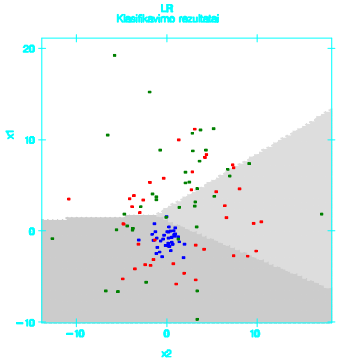
29 pav.Klasifikavimo KDA metodu rezultatai (imtis B120)



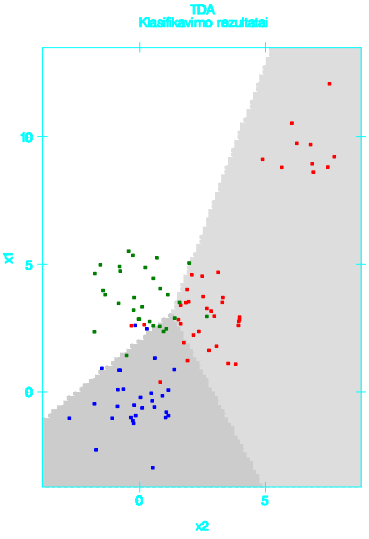
30 pav.Klasifikavimo BDA metodu rezultatai (imtis B120)



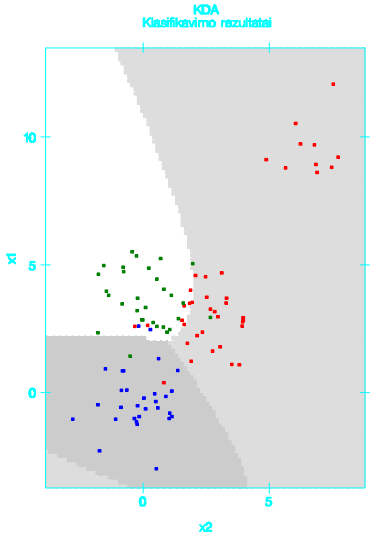
31 pav.Klasifikavimo AKDA metodu rezultatai (imtis B120)



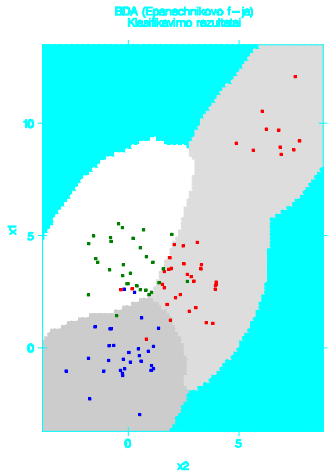
32 pav.Klasifikavimo LR metodu rezultatai (imtis B120)



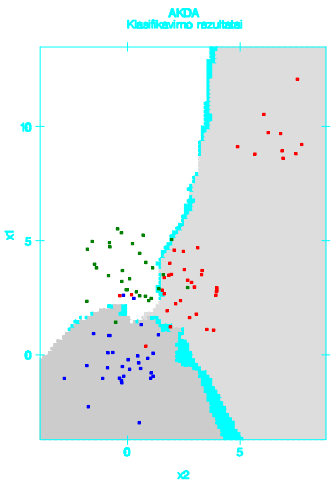
33 pav.Klasifikavimo TDA metodu rezultatai (imtis C130)



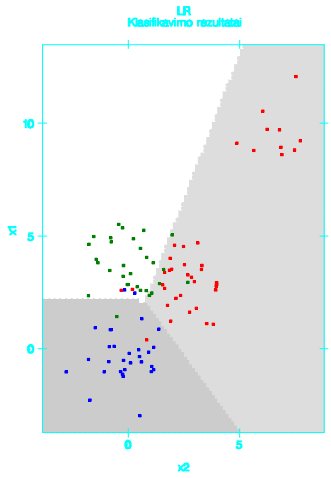
34 pav.Klasifikavimo KDA metodu rezultatai (imtis C130)



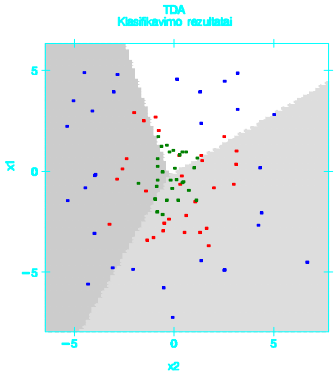
35 pav.Klasifikavimo BDA metodu rezultatai (imtis C130)



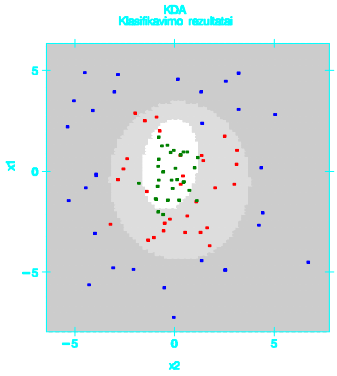
36 pav.Klasifikavimo AKDA metodu rezultatai (imtis C130)



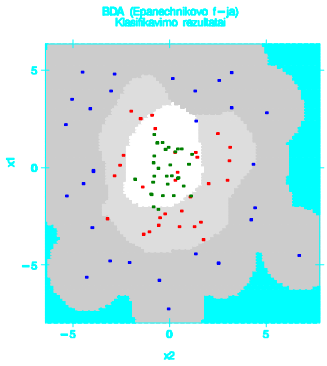
37 pav.Klasifikavimo LR metodu rezultatai (imtis C130)



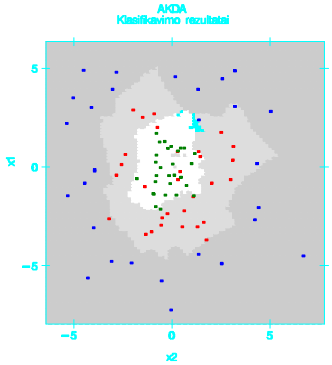
38 pav.Klasifikavimo TDA metodu rezultatai (imtis E120)



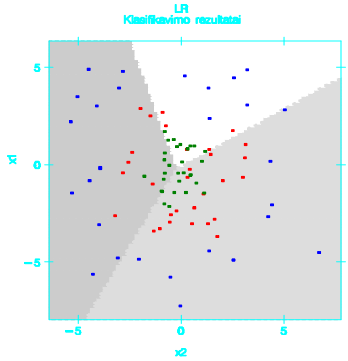
39 pav.Klasifikavimo KDA metodu rezultatai (imtis E120)



40 pav.Klasifikavimo BDA metodu rezultatai (imtis E120)



41 pav.Klasifikavimo AKDA metodu rezultatai (imtis E120)



42 pav.Klasifikavimo LR metodu rezultatai (imtis E120)