VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS INSTITUTAS
PROGRAMŲ SISTEMŲ STUDIJŲ PROGRAMA

Vieno neurono mokymas sprendžiant klasifikavimo uždavinį

2 užduotis

Atliko: 4 kurso 1 grupės studentė
Rosita Raišuotytė

Spalio 08, 2023

TURINYS

1. Įvadas

1.1. Tikslas

Užduoties tikslas - apmokyti vieną neuroną spręsti dviejų klasių uždavinį, ištirti kaip mokymasis priklauso nuo mokymosi grečio, kaip kinta santykinė paklaida bei tikslumas, tyrimą atliekant su dvejomis duomenų aibėmis.

1.2. Uždavinai

- Parsisiųsti ir paruošti duomenis krūties vėžio ir irisų duomenų aibėms.
- Parašyti prorgaminį kodą, kuris įgyvendintų vieno neuorono mokymosi ir testavimo funkcijas.
- Ištirti neurono klasifikavimo tikslumo priklausomybę nuo epochų skaičiaus.
- Ištirti santykinės paklaidos priklausomybę nuo epochų skaičiaus.
- Ištirti kaip mokymosi greitis daro įtaką mokymosi procesui ir rezultatams.

1.3. Sąvokos

Epocha - tai vienas pilnas mokymosi duomenų peržiūros ciklas.

2. Duomenys

Užduotyje naudojami duomenys buvo paimti iš "C Irvine Machine Learning Repository" puslapio. Tyrimui naudoti irisų (angl. iris) ir krūties vėžio (angl. breast cancer) duomenų rinkiniai.

2.1. Irisų duomenų rinkinys

Įrisų duomenų rinkinys turi tris klases, tačiau tyrimui naudoti buvo pasirinktos tik dvi klasės: Versicolor, Virginica. Duomenų rinkinį sudaro:

- 100 duomenų įrašų
- 50 duomenų irašų yra Versicolor klasės
- 50 duomenų irašų yra Virginica klasės
- Versicolor klasė buvo pervadinta į 0
- Virginica klasė buvo pervadinta į 1
- duomenų irašą sudaro 4 požymiai ir klasė

Prieš pradedant dirbti su duomenis, klasės buvo pervadintos į 0 arba 1. Taip pat dėl programinio kodo paprastumo buvo įvestas x_0 požymis, kurio reikšmės visuose įrašuose yra lygios 1.

1 lentelė. Vienas irisų duomenų įrašas prieš ir po pasiruošimo

	x_0	x_1	x_2	x_3	x_4	klasė
prieš	-	7,0	3,2	4,7	1,4	Iris-versicolor
po	1	7,0	3,2	4,7	1,4	0

2.2. Krūties vėžio duomenų rinkinys

Krūties vėžio duomenų rinkinys turi dvi klases: vėžys gali būti nepiktybinis (angl. benign), rinkinyje žymimas 2, ir piktybinis (angl. malignant), žymimas 4. Iš duomenų rinkinio buvo pašalinti nežinomas reikšmes, klaustukus, turintys įrašai. Juos pašalinus duomenų rinkinį sudaro:

- 683 duomenų įrašų
- 444 duomenų įrašai yra 2 klasės
- 239 duomenų įrašai yra 4 klasės
- 2 (nepiktybinis) klasė pervadinta į 0
- 4 (piktybinis) klasė pervadinta į 1
- duomenų irašą sudaro paciento id, 9 požymiai ir klasė

Prieš pradedant dirbti su duomenis, klasės buvo pervadintos į 0 arba 1. Taip pat buvo pašalintas paciento id stulpelis, kadangi jis neteikia jokios informacijos apie krūties vėžį ar jo atsiradimą. Vietoje pašalinto stulpelio dėl programinio kodo paprastumo buvo įvestas x_0 požymis, kurio reikšmės visuose įrašuose yra lygios 1.

2 lentelė. Vienas krūties vėžio duomenų įrašas prieš ir po pasiruošimo

	id	x_0	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	klasė
prieš	1000025	-	5	1	1	1	2	1	3	1	1	2
po	-	1	5	1	1	1	2	1	3	1	1	0

3. Programos kodas

Programos kodą su komentarais galite rasti paspaudę nuorodą čia.

4. Tyrimo variantas

Mano studento pažymėjimo numeris yra 2016026, paskutinis skaitmuo 6. Variantas išsirenkamas pagal paskutinio studento skaitmens dalybos iš 3 liekaną. Mano atveju liekana yra 0. 0 variantas - sigmoidinis neuronas apmokomas naudojant paketinį gradientinį nusileidimą.

4.1. Paketinis gradientinis nusileidimas

Neorono mokymui naudojamas paketinis gradientinis nusileidimas. Gradientinis nusileidimas tai optimizavimo algoritmas, kuris skirtas minimizuoti paklaidos funkcijos reikšmę, kuri matuoja skirtumą tarp prognozuojamos ir tikrosios reikšmės(klasės). Paketinis gradientinis nusileidimas tai toks gradientinio nusileidimo metodas, kurio metu yra iš karto naudojama visas duomenų rinkinys. Skaičiuojamas gradientas kiekvienam įrašui yra sumuojamas ir tik perėjus visą duomenų aibę svoriai yra atnaujinami naudojantis gauto gradiento vidurkiu.

4.2. Sigmoidinis neuronas

Sigmoidinis neuronas - tai neuoronas, kuris naudoja sigmoidinę aktyvacijos funkciją matomą 1 formulėje.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

5. Tyrimas

5.1. Mokymosi ir testavimo duomenys

Duomenys buvo dalinami į mokymosi ir testavimo santykiu 80:20. Prieš dalinant duomenis jie atsitiktinai išmaišomi. Padalinus duomenis klasės tikėtina taip pat pasiskirto santykiu 80:20.

5.1.1. Irisų duomenų rinkinys

Sumaišius ir padalinus irisų duomenų rinkinį gaunama, kad mokymo duomenų rinkinyje yra 80 įrašų, o testavimo 20. Dėl duomenų maišymo klasių pasiskirtymas mokymosi ir testavimo aibėse nevisada būna vienodas, tačiau dažnai tikėtina, kad klasės taip pat pasiskirto panašiu santykiu. Paleidus programinį kodą į ekraną yra išvedama išsami informacija apie duomenų aibes.

5.1.2. Krūvies vėžio duomenų rinkinys

Sumaišius ir padalinus krūties vėžio duomenų rinkinį gaunama, kad mokymo duomenų rinkinyje yra 546 įrašų, o testavimo 137. Dėl duomenų maišymo klasių pasiskirtymas mokymosi ir testavimo aibėse nevisada būna vienodas, tačiau dažnai tikėtina, kad klasės taip pat pasiskirto panašiu santykiu. Paleidus programinį kodą į ekraną yra išvedama išsami informacija apie duomenų aibes.

5.2. Svorių pasirinkimas

Pradiniai svoriai yra generuojami intervale nuo 0 iki 1. Dėl tyrimo tikslumo yra nustatomas pradinis generavimo taškas (angl. seed), kuris kiekvieną kartą sugeneruoja tas pačias svorių reikšmes.

5.3. Tyrimo metodas

Modelis buvo apmokytas su kiekviena duomenų aibe po tris kartus renkantis vis kitą mokymosi greitį. Mokymas vykdomas iki tol kol testavimo aibės klasifikavimo tikslumas pasiekia 0,95 arba kol pereinamos visos epochos.

6. Rezultatai

6.1. Irisų duomenų rinkiniui

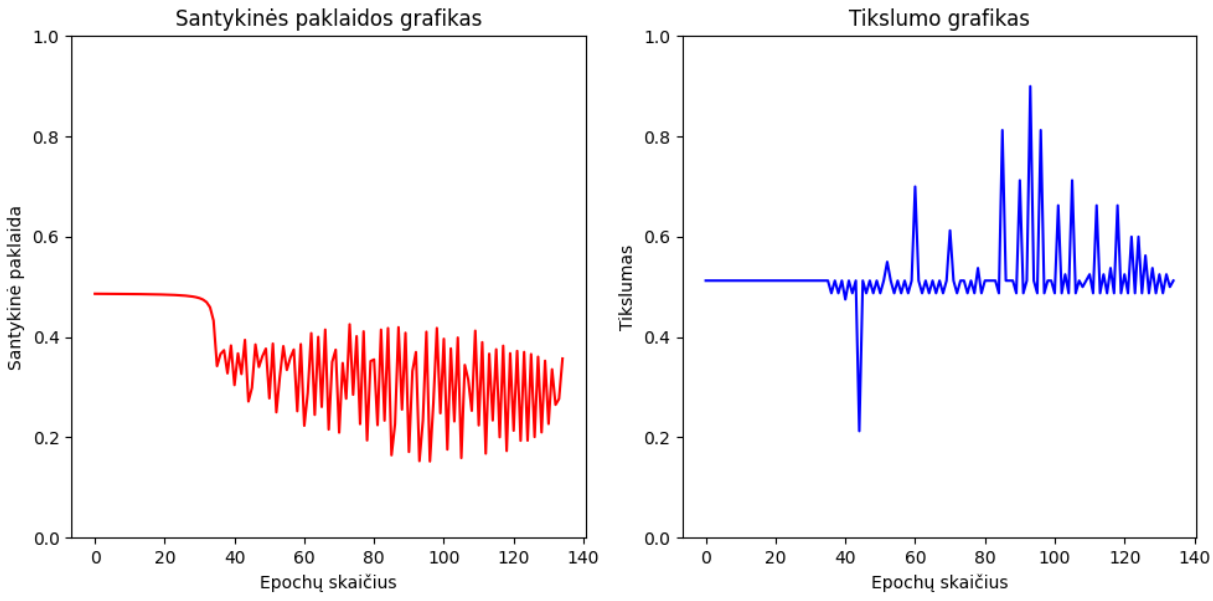
Mokymas buvo vykdomas iki kol pereinamos 300 epochų arba pasiekiamas 0,95 testavimo tikslumas.

6.1.1. Kai mokymosi greitis = 0,9

Modelis, kai mokymosi greitis buvo 0,9, mokėsi 135 epochas, kol testavimo tikslumas pasiekė 0,95. Iš 3 lentelės matyti, kad modelis testavimo metu net lenkė nustatytą reikšmę ir pasiekė 100 procentų tikslumą. Nors paskutinės mokymosi epochos rezultatai gana prasti, testavimo rezultatai puikūs: santykinė paklaida gana maža, o tikslumas labai geras. Žiūrint į testavimo tikslumą galime sakyti, kad ilgiau modelio mokyti nebėra prasmės. Grafikai pavaizduoti 1 paveikslėlyje rodo, kad mokymosi žingsnis buvo per didelis, matoma labai daug šuolių, grafikas nėra tolygus.

3 lentelė. Santykinės paklaidos ir tikslumo rezultatai irisų duomenų rinkiniui, kai mokymosi greitis = 0,9

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,3571	0,1285
Klasifikavimo tikslumas	0,51	1,0



1 pav. Santykinės paklaidos ir tikslumo grafikai irisų duomenų rinkiniui, kai mokymosi greitis = 0,9

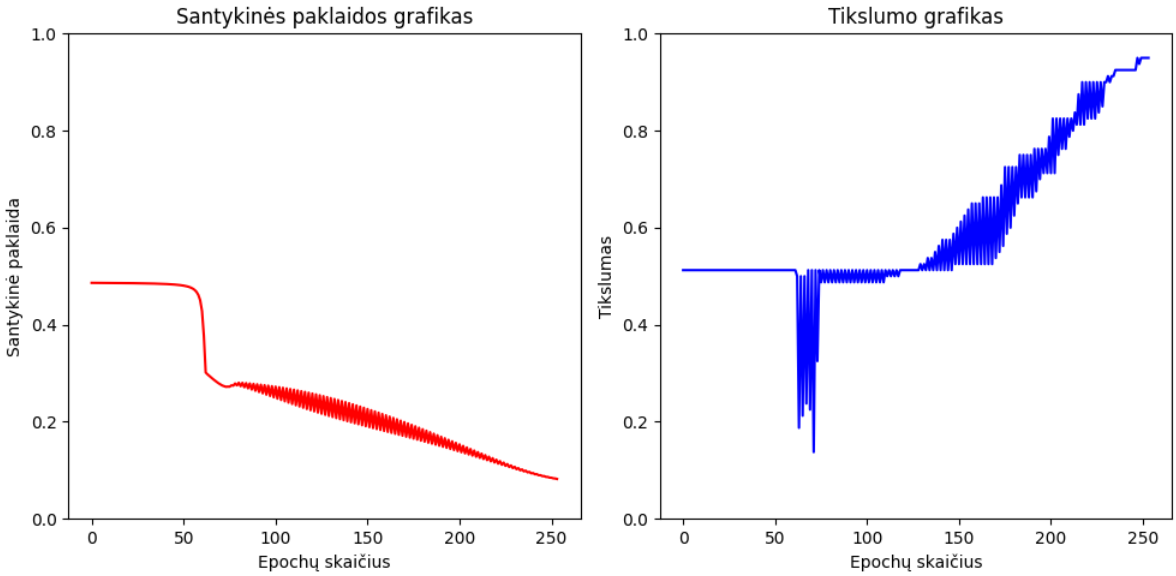
6.1.2. Kai mokymosi greitis = 0,5

Modelis, kai mokymosi greitis buvo 0,5, mokėsi 254 epochas, kol testavimo tikslumas pasiekė 0,95. Iš 4 lentelės matyti, kad paskutinės mokymosi epochos ir testavimo rezultatai labai

panašūs. 2 paveikslėlyje santykinės paklaidos grafike matomi mažesni svyravimai nei tikslumo grafike, tačiau jie yra. Tai mums pasako apie mokymosi greitį, kad jis buvo šiek tiek per didelis. Tolesnis modelio mokymas tikriausiai geresnio tikslumo nepasiektų, nes mokymosi santykinė paklaida jau labai maža.

4 lentelė. Santykinės paklaidos ir tikslumo rezultatai irisų duomenų rinkiniui, kai mokymosi greitis = 0,5

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,0821	0,1207
Klasifikavimo tikslumas	0,95	0,95



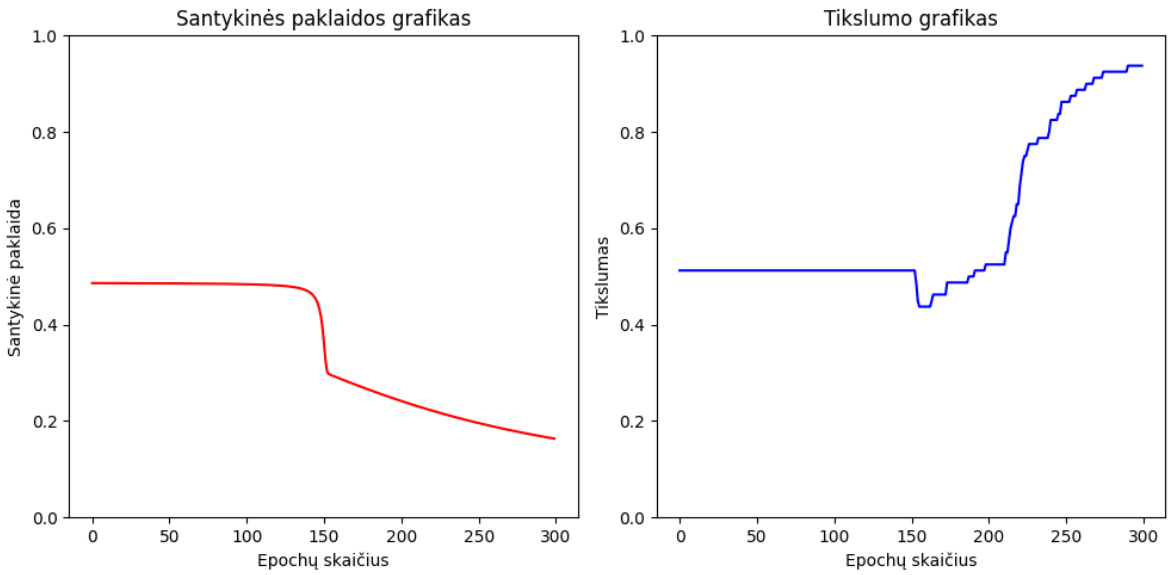
2 pav. Santykinės paklaidos ir tikslumo grafikai irisų duomenų rinkiniui, kai mokymosi greitis = 0,5

6.1.3. Kai mokymosi greitis = 0,2

Modelis, kai mokymosi greitis buvo 0,2, mokėsi visas 300 epochų ir pasiekė 0,95 testavimo tikslumo. 5 lentelėje matomi rezultatai yra geri. Mokymosi ir testavimo santykinės paklaidos rezultatai rodo, kad modelis galėjo dar šiek tiek pasimokyti, tačiau apmokius modelį dar 100 epochų paklaida labai nesumažėjo ir tikslumas nepakilo. Tam įtakos turi gal šiek tiek per mažas mokymosi greitis. 3 paveikslėlio santykinės paklaidos grafike matyti, kad kreivė eina tolygiai, nebėra svyravimų.

5 lentelė. Santykinės paklaidos ir tikslumo rezultatai irisų duomenų rinkiniui, kai mokymosi greitis = 0,2

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,1635	0,1868
Klasifikavimo tikslumas	0,94	0,95



3 pav. Santykinės paklaidos ir tikslumo grafikai irisų duomenų rinkiniui, kai mokymosi greitis = 0,2

6.1.4. Apibendrinimas ir testavimo rezultatai

Irisų duomenų rinkiniui modelis parodė geriausius testavimo rezultatus, kai mokymosi greitis buvo 0,9. Tačiau mokymosi santykinė paklaida dar likusi gana didelė, todėl geriausią modelio apmokymą laikysime antrą bandymą. Antro bandymo, kai mokymosi greitis buvo 0,5, mokymosi paklaida labai sumažėjusi, tai kelia didesnį pasitikėjimą klasifikuojant naujus, nematytus duomenis. Antro bandymo galutiniai modelio svoriai pavaizduoti 6 lentelėje.

6 lentelė. Modelio svorių rinkinys pritaikytas irisų duomenų rinkiniui

x_0	-0,273141
x_1	-1,226686
x_2	-0,797903
x_3	1,652268
x_4	1,465409

Testavimo rezultatai geriausio, antro bandymo metu, kai mokymosi greitis lygus 0,5 matomi 7 lentelėje. Matome, kad buvo sumaišyta tik viena reikšmė - įrašo numeris 20.

7 lentelė. Irisų testinės duomenų aibės tikrosios ir gautos klasifikavimo reikšmes

Nr.	Tikroji reikšmė	Gauta reikšmė
1	0	0
2	1	1
3	1	1
4	0	0
5	1	1
6	1	1
7	0	0
8	1	1
9	0	0
10	0	0
11	0	0
12	0	0
13	0	0
14	1	1
15	1	1
16	1	1
17	1	1
18	0	0
19	0	0
20	0	1

6.2. Krūties vėžio duomenų rinkiniui

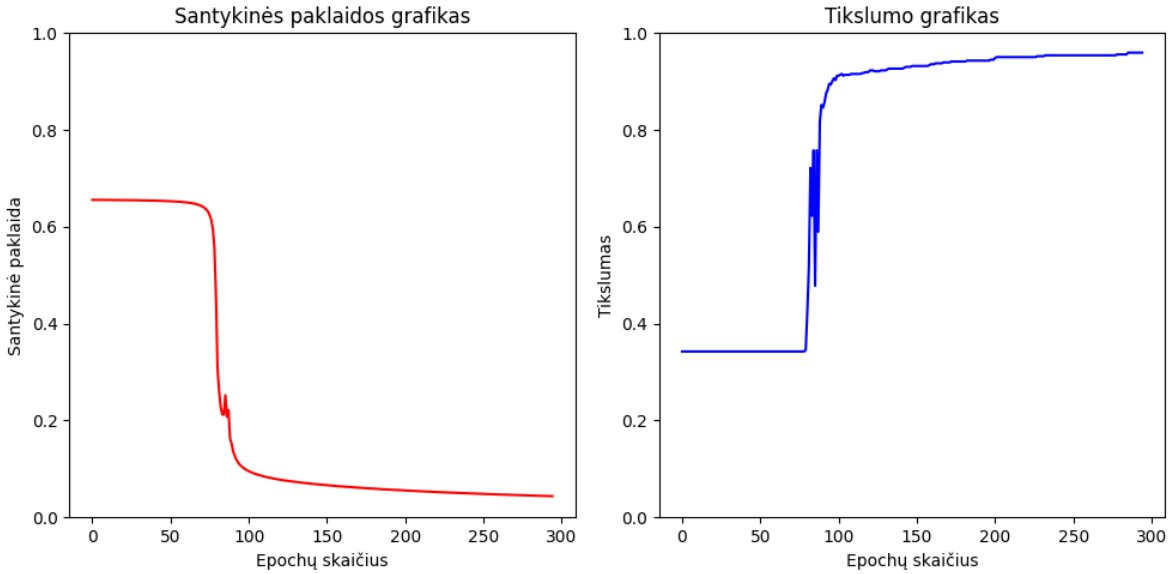
Mokymas buvo vykdomas iki kol pereinamos 500 epochų arba pasiekiamas 0,95 testavimo tikslumas.

6.2.1. Kai mokymosi greitis = 0,9

Modelis, kai mokymosi greitis buvo 0,9, mokėsi 293 epochas, kol testavimo tikslumas pasiekė 0,95. Iš 8 lentelės matyti, kad ir paskutininės mokymosi epochos ir testavimo klasifikavimo tikslumas gavosi didesnis nei norimas pasiekti. Santykinės paklaidos taip pat labai mažos, kad reiškia, kad modelis tikriausiai nebepasiektų geresnių rezultatų. 4 paveikslėlio grafikuose matyti, kad kreivės gana tolygios, vadinasi mokymosi greitis nebuvo pasirinktas per didelis.

8 lentelė. Santykinės paklaidos ir tikslumo rezultatai krūties vėžio duomenų rinkiniui, kai mokymosi greitis = 0,9

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,0437	0,0484
Klasifikavimo tikslumas	0,96	0,96



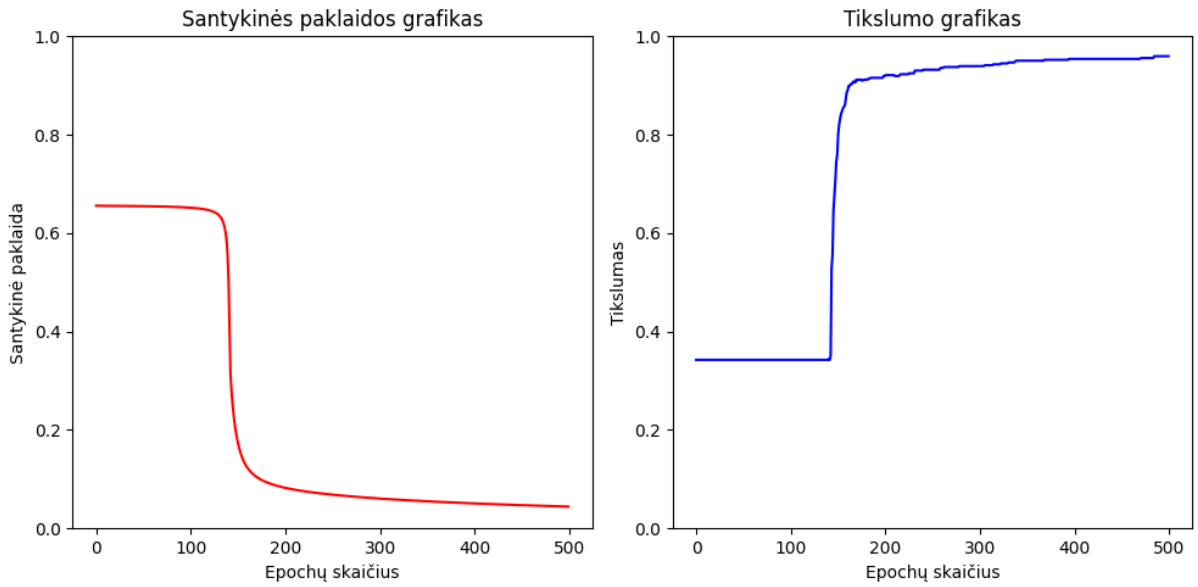
4 pav. Santykinės paklaidos ir tikslumo grafikai krūtues vėžio duomenų rinkiniui, kai mokymosi greitis = 0,9

6.2.2. Kai mokymosi greitis = 0,5

Modelis, kai mokymosi greitis buvo 0,5, mokėsi visas 500 epochų ir pasiekė 0,95 testavimo tikslumą. 9 lentelėje matomi geri mokymosi ir testavimo rezultatai, jie labai panašūs į pirmojo bandymo [8]. Tačiau iš 5 paveikslėlyje esančio santykinės paklaidos grafiko matyti, kad modelis dvigubai ilgiau užtruko kol pradėjo mokytis, vadinasi mookymosi greitis buvo jau šiek tiek per mažas.

9 lentelė. Santykinės paklaidos ir tikslumo rezultatai krūties vėžio duomenų rinkiniui, kai mokymosi greitis = 0,5

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,0442	0,0493
Klasifikavimo tikslumas	0,96	0,95



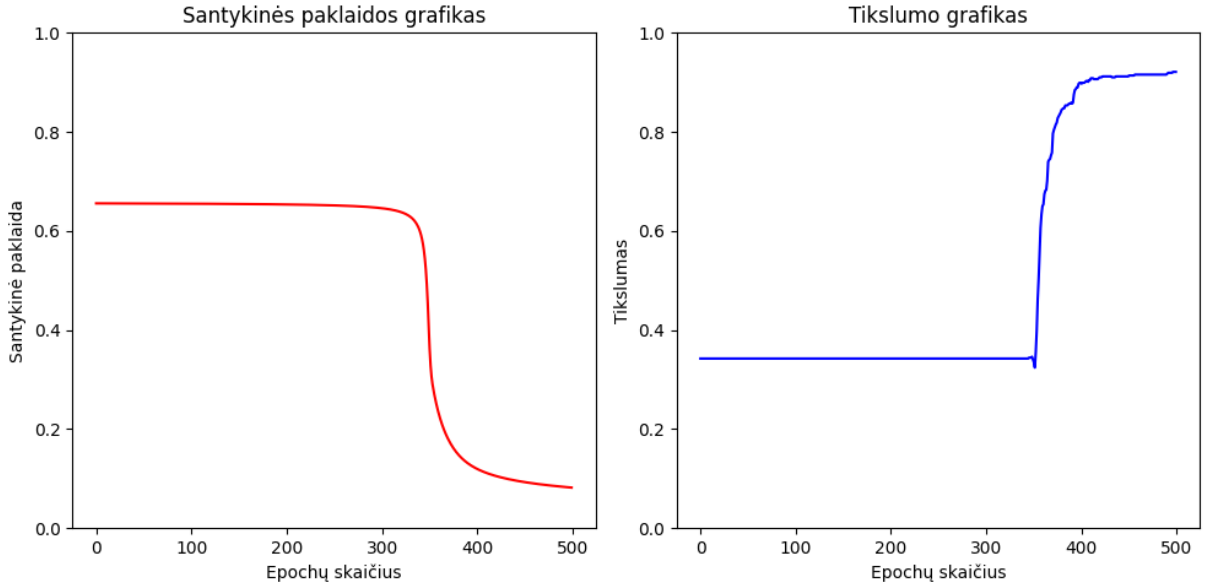
5 pav. Santykinės paklaidos ir tikslumo grafikai krūtues vėžio duomenų rinkiniui, kai mokymosi greitis = 0,5

6.2.3. Kai mokymosi greitis = 0,2

Modelis, kai mokymosi greitis buvo 0,2, mokėsi visas 500 epochų ir nepasiekė norimo testavimo tikslumo. Nors 10 lentelėje matomi rezultatai nėra prasti, tačiau iš 6 paveikslėlyje esančių grafikų matyti, kad mokymasis labai užtruko.

10 lentelė. Santykinės paklaidos ir tikslumo rezultatai krūties vėžio duomenų rinkiniui, kai mokymosi greitis = 0,2

	Paskutinę mokymosi epochą	Testavimo metu
Santykinė paklaida	0,082	0,0925
Klasifikavimo tikslumas	0,92	0,91



6 pav. Santykinės paklaidos ir tikslumo grafikai krūtues vėžio duomenų rinkiniui, kai mokymosi greitis = 0,2

6.2.4. Apibendrinimas ir testavimo rezultatai

Geriausiu bandymu laikysime pirmąjį, nes buvo pasiekti geriausi rezultatai ir grečiausias mokymosi laikas. Nors visų bandymų rezultatai buvo panašūs, mokymosi trukmės labai skyrėsi. Galime teigti, kad krūvies vėžio duomenų rinkiniui modelis greičiausiai ir geriausiai apsimoko, kai mokymosi greitis lygus 0,9. Galutiniai modelio svoriai pavaizduoti 11 lentelėje.

11 lentelė. Modelio svorių rinkinys pritaikytas krūties vėžio duomenų rinkiniui

x_0	-2,288566
x_1	-0,110971
x_2	0,639521
x_3	0,241954
x_4	0,140241
x_5	-0,458733
x_6	0,480063
x_7	-0,281596
x_8	0,276253
x_9	-0,165658

Pirmojo bandymo testavimo rezultatai matomi 12 lentelėje. Suklysta įrašuose: 42, 47, 81, 96, 98, 137. Trejuose įrašuose modelis nepiktybinį naviką pripažino piktybiniu ir likuosiuose trejuose įrašuose atvirkščiai.

12 lentelė. Krūties vėžio testinės duomenų aibės tikrosios ir gautos klasifikavimo reikšmes

Nr.	Tikroji reikšmė	Gauta reikšmė
1	1	1
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	1	1
9	1	1
10	1	1
11	1	1
12	0	0
13	1	1
14	0	0
15	1	1
16	0	0
17	0	0
18	1	1
19	1	1
20	0	0
21	1	1
22	1	1
23	0	0
24	0	0
25	0	0
26	0	0
27	1	1
28	0	0
29	0	0
30	0	0
31	0	0
32	1	1
33	0	0
34	0	0
35	1	1
36	0	0

Continued on next page

lentelė 12 – continued from previous page

Nr.	Tikroji reikšmė	Gauta reikšmė
37	0	0
38	0	0
39	0	0
40	0	0
41	1	1
42	0	1
43	1	1
44	1	1
45	0	0
46	0	0
47	0	1
48	0	0
49	0	0
50	0	0
51	0	0
52	0	0
53	1	1
54	1	1
55	1	1
56	1	1
57	1	1
58	1	1
59	0	0
60	0	0
61	0	0
62	0	0
63	0	0
64	0	0
65	1	1
66	0	0
67	0	0
68	0	0
69	1	1
70	1	1
71	0	0
72	0	0

Continued on next page

lentelė 12 – continued from previous page

Nr.	Tikroji reikšmė	Gauta reikšmė
73	0	0
74	1	1
75	0	0
76	0	0
77	1	1
78	0	0
79	1	1
80	0	0
81	1	0
82	0	0
83	0	0
84	0	0
85	1	1
86	0	0
87	1	1
88	0	0
89	1	1
90	0	0
91	0	0
92	0	0
93	0	0
94	0	0
95	0	0
96	1	0
97	1	1
98	0	1
99	0	0
100	0	0
101	0	0
102	0	0
103	1	1
104	1	1
105	1	1
106	1	1
107	0	0
108	1	1

Continued on next page

lentelė 12 – continued from previous page

Nr.	Tikroji reikšmė	Gauta reikšmė
109	0	0
110	0	0
111	1	1
112	1	1
113	0	0
114	0	0
115	1	1
116	0	0
117	1	1
118	0	0
119	0	0
120	0	0
121	0	0
122	0	0
123	1	1
124	1	1
125	0	0
126	1	1
127	1	1
128	1	1
129	0	0
130	0	0
131	1	1
132	0	0
133	0	0
134	1	1
135	0	0
136	0	0
137	1	0

7. Išvados

Prisiminus uždavinius galime daryti išvadas:

- Mokymo metu santykinė paklaida mažėja didėjant epochų skaičiui, tačiau ar ji mažėja tolygiai arba su šuoliais priklauso nuo mokymosi greičio ir duomenų aibės. Pasiekus momentą, kai modelis nebesimoko, santykinė paklaida nebekinta didėjant epochų skaičiui. Toliau mokyti

neverta.

- Mokymosi metu klasifikavimo tikslumas turėtų didėti didėjant epochų skaičiui, tačiau tai irgi priklauso nuo mokymosi greičio ir duomenų aibės. Taip pat pasiekus tam tikrą tašką, nuo kurio modelis nebesimoko, tikslumas nebedidėja toliau didėjant epochų skaičiui.
- Mokymosi greitis turi didelę įtaką mokymosi procesui ir rezultatams, nuo jo priklauso kaip greitai mes pasieksime norimus rezultatus ir ar iš viso juos galime pasiekti. Taip pat nuo jo priklauso ir mokymosi trukmė. Per didelis mokymosi greitis gali įtakoti modelį mokytis šuoliais, nesuteikiant garantijos, kad bus pasiektas maximalus rezultatas.