

I. Tuleiko, M. Gasparoviča (zinātniskā vadītāja)

PIEDERĪBAS FUNKCIJU KONSTRUĒŠANAS METOŽU IZPĒTE BIOINFORMĀTIKAS DATU ANALĪZĒ

Mūsdienās ļoti aktuāla ir izplūdušī loģika, izplūdušās datu kopas un to apstrādes jēdzieni, jo izplūdušī loģika ir daudz tuvāka cilvēka domāšanai kā binārie jēdzieni. Bioinformātika ir zinātne, kur izplūdušī loģika pakāpeniski tiek pielietota arvien vairāk, ņemot vērā bioinformātikas datu īpatnības – tie ir izplūduši, skaitliski nepārtraukti vai kategoriski, tiem ir parasti nelielas datu kopas un ļoti liels atribūtu skaits.

Darba mērķis ir atrast piemērotāko piederības funkciju konstruēšanas metodi bioinformātikas datu kopai. Līdz ar to tika veikts pētījums par piederības funkciju konstruēšanas metodēm, ko iespējams izmantot arī bioinformātikas datiem. Tika noteikta katras metodes piemērotība izvēlētajai datu kopai un tika veikta piederības funkciju konstruēšanas metožu salīdzinošā analīze, izmantojot reālas datu kopas. Pētījumā apskatītas četras piederības funkciju konstruēšanas metodes, to salīdzinājums apskatāms 1. tabulā.

1. tabula

Piederības funkciju konstruēšanas metodes

Nr	Metodes nosaukums	Īss apraksts
1.	Piederības funkciju konstruēšanas metode, balstīta uz ekspertu vērtējumiem	Katra atribūta apmācības kopas vērtības tiek sadalītas vairākos intervālos. Katra atribūta katrai vērtībai eksperts nosaka piederības funkcija katram no intervāliem. Tā kā bioinformātikas eksperts nebija pieejams, tas tika imitēts ar algoritma palīdzību, kas ģenerēja piederības funkcijas atbilstoši vērtības atrašanās vietai atribūta vērtību kopā. Piederības funkciju konstruēšanas metode, kas ir vissubjektīvākā.
2.	Piederības funkciju konstruēšanas metode, balstīta uz pāru salīdzinājumiem	Katra atribūta apmācības kopas vērtības tiek sadalītas vairākos intervālos un piederības funkcijas tiek konstruētas, izmantojot Sāta skalu un izveidojot intervālu salīdzinājuma matricu. Šīs matricas vektors veido piederības funkciju. Metode izmanto eksperta salīdzinājumu, tomēr ir mazāk atkarīga no cilvēciskā faktora, kā pirmā apskatītā metode.
3.	Lingvistisku jēdzienu piederības funkciju konstruēšanas metode, izmantojot statistiskus datus	Metode balstīta uz vairāku ekspertu novērojumiem par notikuma biežumu intervālos.
4.	Piederības funkciju konstruēšanas parametriskā pieeja	Metode balstīta uz jaunu piederības funkciju konstruēšanu, izmantojot jau zināmās piederības funkcijas.

Lai salīdzinātu piederības funkcijas, tika izmantotas reālas datu kopas ar bioinformātikas datiem, to apraksts redzams 2. tabulā.

Datu kopa tika sadalīta apmācības un testa kopās, izmantojot nejaušības principu, kur 70% ierakstu tika iekļauti apmācības kopā un 30% ierakstu – testa kopā.

Izmantotās datu kopas

Nr	Datu bāze	Ieraksti	Atribūti	Klases	Apraksts
1.	Gastrīta audzēju dati	328	1229	2 (GaCa, HD)	Normalizēta datu kopa, daudz iztrūkstošo vērtību, atribūti šifrēti, nav zināmi gēnu patiesie nosaukumi.
2.	Leikēmijas dati (broadinstitute.org)	29	22283	2 (CS, CR)	Nenormalizēta datu kopa, minimāli iztrūkstoši dati.

Izmantojot dažādās piederības funkciju konstruēšanas metodes iegūtas piederības funkcijas, veikta likumu veidošana un testa kopas klasificēšana, izmantojot FuzzyPrism algoritmu. Piederības funkciju konstruēšanas metodes, balstītas uz ekspertu vērtējumiem un Piederības funkciju konstruēšanas metodes, balstītas uz pāru salīdzinājumiem pretstatījums atbilstoši iegūtajiem rezultātiem pirmajai datu kopai redzams 3. tabulā.

Metožu salīdzinājums

	Pareizi klasificēti	Nepareizi klasificēti	Netika klasificēti
Piederības funkciju konstruēšanas metode, balstīta uz ekspertu vērtējumiem	23	20	56
	Metodei nav pastiprinātas sarežģītības, ļoti nozīmīgs ir laba eksperta viedoklis, kas spēcīgi ietekmē metodes lietderību.		
Piederības funkciju konstruēšanas metode, balstīta uz pāru salīdzinājumiem	73	11	15
	Šī metode uzrāda labākus rezultātus kā pirmā metode, tomēr nepieciešami papildus eksperimenti, lai to apstiprinātu. Metode ir sarežģīta un laikietilpīga aprēķinu algoritma izveidošanas ziņā.		

Perspektīvā ir paredzēts veikt visu metožu salīdzinājumu abām datu kopām, kas ir apjomīgs darbs, ņemot vērā datu kopu specifiku. Tomēr no iegūtajiem rezultātiem iespējams secināt, ka izmantojot atbilstošāko piederības funkciju konstruēšanas metodi, testa datu kopas ierakstu klasifikācija uzlabojas vairākas reizes.