M. Gasparoviča, L.Aleksejeva (zinātniskā vadītāja)

IZPLŪDUŠO ALGORITMU PIELIETOŠANA KLASIFIKĀCIJAS UZDEVUMU RISINĀŠANĀ

Pētījuma aktualitāte ir saistīta ar cilvēka spēju darīt zināmu savu attieksmi pret vecumu, augumu vai jebkuru lingvistisku jēdzienu datoram saprotamā, interpretējamā un apstrādājamā veidā. Ārkārtīgi reti reālajā dzīvē ir sastopami klasiski dati, kur viegli noteikt to piederību noteiktai klasei, jo ir iespēja piederēt klasei tikai daļēji. Tādējādi arī algoritmi, kas spēj klasificēt izplūdušus datus, pakāpeniski kļūst aizvien vajadzīgāki un neaizvietojamāki dažādās dzīves jomās. Pētījuma objekts ir datu ieguvē izmantojamo izplūdušo klasifikācijas algoritmu iespēju pētīšana un analīze. Pētījuma mērķis ir izpētīt un analizēt izplūdušos klasifikācijas algoritmus, sniegt rekomendācijas algoritmu praktiskam pielietojumam.

Darbā tiek izpētīti un pielietoti divi izplūdušie klasifikācijas algoritmi – nozīmīgo atribūtu un piederības funkciju meklēšanas algoritms (NAPFM algoritms) un izplūdušais PRISM algoritms, kā arī pētītas radniecīgas metodes izplūdušu datu klasificēšanā. Pētīto algoritmu salīdzinājums aplūkojams 1. tabulā.

Algoritmu salīdzinājums 1. tabula

Algoritma nosaukums	**Darbības princips**	**Prasības datu kopām**	**Priekšrocības**
NAPFM algoritms	Meklē nozīmīgāko (vai vairākus) atribūtu, no kura veido likumus, pārējos atribūtus atmetot	Skaitliski, nepārtraukti	*Automātiska piederības funkciju konstruēšana; *Neliels lēmumu likumu skaits.
Izplūdušais PRISM	Meklē labākās atribūtu – vērtību izplūdušās piederības funkcijas	Kategoriski dati	*Universālāks, kategoriski dati; *Iekļauj eksperta viedokli.

Pētījumā ar algoritmiem noskaidrojās, ka nozīmīgo atribūtu un piederības funkciju meklēšanas algoritmam ir visaugstākās prasības pret izmantojamās datu kopas parametriem, tāpēc tika meklētas datu kopas, kas atbilstu tā prasībām. Izskatot UCI datu repozitorijā pieejamās 198 datu bāzes, tika atlasītas 10 (skat. 2. tabulu), kas tika pārbaudītas darbā ar algoritmu un no tām atlasītas tikai dažas perspektīvākās tālākai izmantošanai.

Pārbaudītās datu kopas 2. tabula

Nosaukums no UCI	**Ieraksti**	**Atribūti**	**1.nozīmīg.**	**2.nozīmīg.**	**Derīguma pakāpe <0.1**	**Perspektīva**
Blood Transfusion	748	5	0.08	0.033	0.89	Nav
Pima Indians Diabetes	768	8	0.29	0.23	0.55	Nav
Ecoli	336	8	0.28	0.18	0.59	Nav
Gamma Telescope	19020	11	1	1	0	Ir
Horse Colic	368	27	0.34	0.28	0.48	Nav
Forest Fires	517	13	0.09	0.008	0.903	Nav
Iris	150	4	0.78	0.7	0.066	Ir
Ionosphere	351	34	0.63	0.24	0.19	Ir/Nav
Haberman's Survival	306	6	0.062	0.045	0.89	Nav
Auto MPG	392	8	0.97	0.62	0.03	Ir

Kā redzams rezultātu 3. tabulā, tad visaugstāko rezultātu uzrāda izplūdušais PRISM algoritms, ar sīkāku atribūtu sadalījumu intervālos. Tomēr arī nozīmīgo atribūtu un piederības funkciju meklēšanas algoritms uzrāda augstas precizitātes rezultātus. Labākus rezultātus iespējams iegūt, ja nepārtraukta datu kopa tiek pārveidota kategoriskā, cenšoties palielināt jauno vērtību skaitu, taču šeit jāievēro princips, lai ieguldītais darbs būtu adekvāts iegūtajam rezultāta uzlabojumam.

Īrisa ziedu datu kopa 3. tabula

	Apmācības kopa	Testa kopa	Kļūdaini klasificētie	Precizitāte	Kļūda	Komentāri
NAPFM algoritms	105	35	2	0,94	0,06	Klasiskais sadalījums (70:30)
	102(96)	48(54)	0	0,96	0,04	Trīskārtīgā šķērsvalidācija 150 ier.
	75	75	4	0,96	0,04	Apmācība: testēšana (50:50)
	135	15	1	0,93	0,07	Lielāka apmācības, mazāka testa kopa
	105	15	1	0,93	0,07	Atmesti trīsdesmit ieraksti
	78(81)	42(39)	1	0,96	0,04	Trīskārtīgā šķērsvalidācija 120 ier.
Izplūd. PRISM	105	45	0	1,00	0,0	Klasiskais sadalījums (līdz 6 interv.)
	105	45	3	0,93	0,07	Klasiskais sadalījums (3 interv.)
	96(102)	54(48)	4	0,88	0,12	Trīskārtīgā šķērsvalidācija (3 interv.)

Perspektīvi būtu uzlabot nozīmīgo atribūtu un piederības funkciju meklēšanas algoritmu, lai tas spētu darboties arī ar kategoriskiem datiem, jo šis algoritms mazāk ietekmējas no subjektīvas informācijas. Taču izplūdušais PRISM algoritms ir universālāks un to iespējams izmantot dažādu datu kopu klasificēšanā, kā arī dažādu reālu klasifikācijas problēmu risināšanā.