REPUBLIKA E SHQIPËRISË

UNIVERSITETI I TIRANËS

FAKULTETI I SHKENCAVE TË NATYRËS

DEPARTAMENTI I INFORMATIKËS

ERIND BEDALLI

DISERTACION

LOGJIKA E TURBULLT DHE APLIKIME TË SAJ NË MODELIM

Udhëheqës Shkencor

PROF. DR. ILIA NINKA

TIRANË, 2014

UNIVERSITETI I TIRANËS

FAKULTETI I SHKENCAVE TË NATYRËS

DEPARTAMENT I INFORMATIKËS

DISERTACION

I PARAQITUR NGA:

M. Sc. ERIND BEDALLI

PËR MARRJEN E GRADËS SHKENCORE

DOKTOR

SPECIALITETI: INFORMATIKË

TEMA

"LOGJIKA E TUBULLT DHE APLIKIME TË SAJ NË MODELIM"

Mbrohet në datën ____.____._____ para jurisë:

1. _____ Kryetar
2. _____ Anëtar (Oponent)
3. _____ Anëtar (Oponent)
4. _____ Anëtar
5. _____ Anëtar

Përmbledhje

Logjika e turbullt është një përgjithësim i logjikës klasike duke prezantuar konceptin e vërtetësisë së pjesshme. Ndërkohë që vlera e vërtetësisë e një pohimi në logjikën klasike mund të jetë ose 1 (i vërtetë) ose 0 (i gabuar), në logjikën e turbullt shkalla e vërtetësisë e një pohimi mund të jetë çdo vlerë në segmentin e vazhduar [0,1]. Ky ndryshim i "vogël" në përkufizim sjell një zhvillim të madh në aftësinë e logjikës së turbullt për të paraqitur dhe trajtuar forma të ndryshme të informacionit dhe dijes.

Në këtë disertacion fillimisht do të trajtohen themelet teorike të logjikës së turbullt, si bashkësitë e turbullta, shkalla e anëtarësisë, relacionet e turbullta, operatorët, procedurat e shturbullimit si dhe zgjatime të logjikës së turbullt si logjika e turbullt intuitive dhe logjika e turbullt e llojit të dytë.

Më tej do të studiohen aplikime të saj në modelim, për të eksploruar strukturën e të dhënave, për të kuptuar dhe interpretuar modelet në të dhëna dhe për të krijuar modele të reja prej modeleve ekzistuese. Në këtë pjesë kemi zhvilluar hulumtime të thelluara në një nga komponentët më të rëndësishëm të mësimit pa mbikëqyrje: fusha e analizës së turbullt të grupimeve (particionale dhe hierarkike). Ndër metodat e grupimeve particionale kemi studiuar, eksperimentuar dhe kemi propozuar teknika sintonizimi dhe hibridizimi në lidhje me algoritmin e grupimit të turbullt të mesoreve dhe një sërë variantesh të tij si algoritmet Gustafson-Kessel,Gath-Geva, grupimi i turbullt i mesoreve bazuar në bërthama etj.

Gjithashtu kemi kryer një studim teorik si dhe implementimin të shoqëruar me procedura eksperimentale të teknikave të përshtatjes së algoritmeve të analizës së turbullt të grupimeve për të vepruar në kushte të mësimit gjysmë të mbikëqyrur si dhe në përdorimin e metodave ansambël (homogjene dhe heterogjene) në grupime për të ndërtuar modele grupues me saktësi dhe stabilitet më të lartë.

Si rezultat kryesor i këtij studimi ofrohet një platformë e pasur mjetesh dhe metodash të bazuara në logjikën e turbullt që mundësojnë eksplorimin e modeleve të "fshehura" në depozita të mëdha të dhënash, ndërtimin e modeleve të bazuar në rregulla që kryejnë përcaktimet sasiore dhe zbatimin e këtyre teknologjive në projektimin, zhvillimin, testimin dhe verifikimin modeleve të jetës reale.

Fjalë kyçe: *logjika e turbullt, analiza e turbullt e grupimeve, algoritmi i turbullt i mesoreve, verifikimi i grupimeve, mësimi gjysmë-i mbikëqyrur, metodat ansambël*

Abstract

Fuzzy logic is a generalization of the classical logic introducing the concept of partial degree of truth. While the truth degree of a proposition in the classical logic is either 1 (true) or 0 (false), the degree of truth of a proposition in fuzzy logic can be any value in the continuous segment [0,1]. This "slight" change in the definition implies a great improvement in the capability of fuzzy logic to represent and treat various forms of information and knowledge. In this dissertation firstly will be discussed the theoretical foundations of fuzzy logic like fuzzy sets, membership degree, fuzzy relations, operators, defuzzification procedures and extensions of fuzzy logic like the intuitionistic fuzzy logic and type 2 fuzzy logic.

Further we will study applications of fuzzy logic in modeling to explore structure of data, to comprehend and interpret models in the data and to create new models from the existing ones. In this part we have conducted detailed research in one of the most important components of unsupervised learning: the field of fuzzy cluster analysis (both partitional and hierarchical). Among the partitional fuzzy clustering methods we have studied, experimented and proposed new tuning and hybridization techniques for the fuzzy c-means algorithm and several variations of it like the Gustafson-Kessel and Gath-Geva algorithms, kernel-based fuzzy c-means algorithm etc.

Also we have performed a theoretical study associated with implementation and experimental procedures about the techniques of adapting the fuzzy cluster analysis algorithms to operate under the conditions of semi-supervised learning. Moreover we have theoretically and experimentally researched the usage of cluster ensemble methods (homogeneous and heterogeneous) in clustering to construct clustering models of improved accuracy and stability.

As the main result of this study, a rich platform of instruments and methods based in fuzzy logic is offered to enable the exploration of the "hidden" models in large deposits of data, the construction of models based in rules which perform quantitative evaluations about the data and the application of these technologies in the design, development, testing and verification of real life models.

***Keywords**: fuzzy logic, fuzzy cluster analysis, fuzzy c-means algorithm, cluster validation, semi-supervised learning, ensemble methods*

Përmbajtja e lëndës

Përmbledhje iii
Abstract iv
Lista e figurave viii
Lista e tabelave x
1. HYRJE 1
1.1 Korniza teorike dhe motivimi 1
1.2 Qëllimi 3
1.3 Vështrim i përgjithshëm i përmbajtjes 3
2. LOGJIKA E TURBULLT 5
2.1 Bashkësitë e turbullta 6
2.1.1 Funksionet e anëtarësisë 9
2.1.2 Veprimet mbi bashkësitë e turbullta dhe vetitë e tyre 12
2.1.3 Transformimi në bashkësi klasike 16
2.2 Relacionet e turbullta 17
2.3 Implikimet e turbullta 19
2.4 Zgjatime të logjikës së turbullt 21
2.4.1 Logjika e turbullt intuitive 21
2.4.2 Logjika e turbullt e llojit 2 23
3. ANALIZA E GRUPIMEVE 25
3.1 Klasifikimi i algoritmeve grupues 27
3.2 Funksionet e distancës dhe ngjashmërisë 28
3.3 Algoritmet hierarkikë 30
3.3.1 Algoritmet hierarkikë bashkues 30
3.3.2 Algoritmet hierarkikë ndarës 33
3.4 Algoritmet particionalë 34
3.4.1 Algoritmet particionalë klasikë (të fortë) 34

3.4.1.1 Algoritmi klasik i mesoreve 35
3.4.2 Algoritmet particionalë të turbullt (të butë) 36
3.4.2.1 Algoritmi i turbullt i mesoreve 37
3.4.2.2 Algoritmi Gustafson-Kessel 38
3.4.2.3 Algoritmi Gath-Geva 39
3.4.2.4 Algoritmi i grupimit të turbullt bazuar në bërthama 41
3.4.2.5 Algoritmi posibilistik i mesoreve 42
4. TEKNIKA SINTONIZUESE DHE VLERËSUESE, KRAHASIME EKSPERIMENTALE PËR ALGORITMET E GRUPIMIT. 45
4.1 Bashkësitë e të dhënave të përdorura në eksperimente 46
4.2 Procedura eksperimentale sintonizuese dhe krahasuese 48
4.2.1 Eksperimente sintonizuese rreth distancave Minkovski 48
4.2.2 Eksperimente sintonizuese rreth koeficientit të turbullsisë 51
4.2.3 Eksperimente krahasuese rreth performancës së algoritmeve të grupimit 54
4.3 Verifikimi i grupimeve 56
4.3.1 Indekset e verifikimit 58
4.3.2 Rezultate ekperimentale mbi indekset e verifikimit 60
4.4 Analiza e grupimeve në eksplorimin e të dhënave të një sistemi arsimor 62
4.4.1 Metodologjia e zbatuar 63
4.4.2 Krijimi i profileve karakterizues të performancës së përgjithshme 64
4.4.3 Krijimi i profileve karakterizues të performancës specifike 65
5. PËRSHTATJA E ALGORIMIT TË TURBULLT TË MESOREVE PËR NJË PËRQASJE TË MËSIMIT GJYSMË-TË MBIKËQYRUR 68
5.1 Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur i bazuar në të dhëna të kategorizuara 69
5.2 Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur i bazuar në kufizime 72
5.3 Studime eksperimentale mbi algoritmet e grupimit të turbullt gjysmë-të mbikëqyrur 74

5.3.1 Eksperimente mbi algoritmin e turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në të dhëna të kategorizuara 74
5.3.2 Eksperimente mbi algoritmin e turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në kufizime 76
6. METODAT ANSAMBËL NË ANALIZËN E TURBULLT TË GRUPIMEVE 78
6.1 Parimet e përgjithshme të projektimit të metodave ansambël 79
6.2 Një model homogjen i një metode ansambël të grupimit të turbullt 81
6.3 Një model heterogjen i një metode ansambël të grupimit të turbullt 83
6.4 Studime eksperimentale mbi metodat ansambël 84
6.4.1 Eksperimente mbi modelin homogjen të metodave ansambël 84
6.4.2 Eksperimente mbi modelin heterogjen të metodave ansambël 85
7. PËRFUNDIME, KONTRIBUTE DHE DREJTIME PUNE KËRKIMORE NË TË ARDHMEN 87
7.1 Përfundime 87
7.2 Kontribute 88
7.3 Drejtime për zhvillim pune kërkimore në të ardhmen 89
Referenca 91

Lista e figurave

Figurë 1 - Funksioni i anëtarësisë për bashkësinë e turbullt diskrete A_1 7
Figurë 2 - Funks ioni i anëtarësisë për bashkësinë klasike diskrete A_2 7
Figurë 3 - Funksioni i anëtarësisë për bashkësinë e turbullt të vazhdueshme A_3 7
Figurë 4 - Funksioni i anëtarësisë për bashkësinë klasike të vazhdueshme A_4 7
Figurë 5 - Bashkësia e njerëzve të gjatë si bashkësi klasike 8
Figurë 6 - Bashkësia e njerëzve të gjatë si bashkësi e turbullt 8
Figurë 7 - Karakteristikat e funksionit të anëtarësisë 10
Figurë 8 - Disa nga format kryesore të funksioneve të anëtarësisë 12
Figurë 9 - Veprimet themelore mbi bashkësitë e turbullta 13
Figurë 10 - Transformimi në bashkësi klasike me anë të prerjes α 16
Figurë 11 - Mënyra të paraqitjes së relacioneve të turbullta 18
Figurë 12 - Shembull i zonës së anëtarësisë për një bashkësi të turbullt të llojit 2 23
Figurë 13 – Etapat kryesore të një procesi të analizës së grupimeve 26
Figurë 14 – Pseusdokodi përshkrues i algoritmeve hierarkikë bashkues 31
Figurë 15 - Dendrograma përfundimtare e gjeneruar nga algoritmi hierarkik i lidhjes së njëfishtë 33
Figurë 16 – Shembuj rezultatesh të algoritmeve hierarkikë mbi bashkësi relativisht të mëdha të dhënash të paraqitura vizualisht (me anë të dendrogrameve) 33
Figurë 17 - Pseudokodi përshkrues i algoritmit të fortë të mesoreve 36
Figurë 18 - Pseudokodi përshkrues i algoritmit të turbullt të mesoreve 38
Figurë 19 - Pseudokodi përshkrues i algoritmit Gustafson-Kessel 39
Figurë 20 - Pseudokodi përshkrues i algoritmit Gath-Geva 40
Figurë 21- Pseudokodi përshkrues i algoritmit të grupimit të turbullt të bazuar në bërthama 42
Figurë 22 - Rrethana ku shpërndarja probabilistike nuk do ishte e përshtatshme për elementin x_2 42
Figurë 23 - Pseudokodi përshkrues i algoritmit posibilistik të mesoreve 44
Figurë 24 - Shkalla e saktësisë në varësi të rendit të distancës Minkovski për disa bashkësi të dhënash 50
Figurë 25 - Shkalla e saktësisë në varësi të koeficientit të turbullsisë për disa bashkësi të dhënash 53
Figurë 26 - Numri i iteracioneve të algoritmeve për disa bashkësi të dhënash 55
Figurë 27 - Koha e ekzekutimit të algoritmeve për disa bashkësi të dhënash 56

Figurë 28 - Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur bazuar në të dhëna të kategorizuara 70
Figurë 29 - Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur bazuar në kufizime 73
Figurë 30- Shkalla e saktësisë së algoritmit në varësi të përqindjes së elementëve të kategorizuar 75
Figurë 31 - Shkalla e saktësisë së algoritmit në varësi të numrit të kufizimeve 77
Figurë 32 - Pseudokodi përshkrues i një modeli homogjen të një metode ansambël të grupimit të turbullt 82
Figurë 33- Pseudokodi përshkrues i një modeli heterogjen të një metode ansambël të grupimit të turbullt 84

Lista e tabelave

Tabelë 1 - Vetitë e veprimeve në bashkësitë e turbullta 14
Tabelë 2 - Veprimet themelore mbi relacionet e turbullta 19
Tabelë 3 - Format kryesore të funksioneve implikim të turbullta 21
Tabelë 4 - Detajet e bashkësive të të dhënave të përdorura 47
Tabelë 5 - Shkalla e saktësisë në ekperimentet përkatëse në lidhje me distancën Minkovski 49
Tabelë 6 - Shkalla e saktësisë në ekperimentet përkatëse në lidhje me koeficientin e turbullsisë 52
Tabelë 7- Numri mesatar i përsëritjeve (iteracioneve) për algoritmet përkatëse 55
Tabelë 8 - Koha mesatare e ekzekutimit (në sekonda) për algoritmet përkatëse 55
Tabelë 9 - Rezultate eksperimentale të indekseve të verifikimit mbi bashkësinë Synth1 . 60
Tabelë 10 -Rezultate eksperimentale të indekseve të verifikimit mbi bashkësinë Synth2 61
Tabelë 11 - Përmbledhje e procedurës verifikuese me numrin optimal të grupimeve për secilin rast 61
Tabelë 12 - Vlerat e anëtarësisë së disa studentëve në grupimet e gjeneruara 65
Tabelë 13 - Vlerat e anëtarësisë për disa studentë në kategoritë e gruplëndës së parë 66
Tabelë 14 - Vlerat e anëtarësisë për disa studentë në kategoritë e gruplëndës së dytë 66
Tabelë 15- Shkalla e saktësisë së algoritmit në varësi të përqindjes së elementëve të kategorizuar 75
Tabelë 16 - Shkalla e saktësisë së algoritmit në varësi të numrit të kufizimeve 76
Tabelë 17 - Shkalla e sakësisë në modelin homogjen të metodave ansambël 85
Tabelë 18 - Shkalla e sakësisë në modelin heterogjen të metodave ansambël 86

1. HYRJE

Logjika e turbullt është një nga fushat e studimit që në dekadat e fundit ka tërhequr masivisht interesin e komunitetit të kërkimit shkencor. Ajo është një përgjithësim i logjikës klasike duke zhvilluar konceptin e vërtetësisë së pjesshme. Ky modifikim në konceptimin e vërtetësisë e ka bërë logjikën e turbullt një mjet të përshtatshëm në trajtimin e disa formave të pasigurisë, si rrjedhim ajo ka gjetur zbatime të suksesshme në një gamë të gjerë disiplinash si inxhinieri, teori kontrolli, inteligjencë artificiale, shkenca konjitive, bioinformatikë, mjekësi, ekonomi etj [1].
Një nga zbatimet më të rëndësishme të logjikës së turbullt është në lidhje me problemin e analizës së të dhënave, ku shfrytëzohet në etapën e modelimit për të eksploruar strukturën e të dhënave, për të kuptuar dhe interpretuar modelet në të dhëna dhe për të krijuar modele të reja prej modeleve ekzistuese [2]. Pikërisht në këtë drejtim është orientuar dhe punimi ynë. Në këtë kapitull prezantues do të paraqesim kornizën e përgjithshme teorike, motivimin për realizimin e këtij punimi, një përshkrim të qëllimeve të këtij punimi dhe një përshkrim të organizimit të punimit në kapitujt përkatës.

1.1 Korniza teorike dhe motivimi

Një nga format kryesore të analizës së të dhënave, pjesë e mësimit pa mbikëqyrje, është analiza e grupimeve. Në procesin e analizës së grupimeve të dhënat synohen të ndahen në grupime (klasa), ku të dhënat të jenë sa më të ngjashme me të dhënat brenda të njëjtit grupim dhe sa më të ndryshme me të dhënat në grupimet e tjera. Në këtë kuadër ngjashmëria konceptohet si një madhësi matematikisht e përkufizueshme dhe ekzistojnë një sërë mënyrash për ta shprehur atë, siç do ta konkretizojmë në kapitujt vijues. Analiza e grupimeve përfshihet në mësimin pa mbikëqyrje duke qenë se operohet pa ndonjë informacion të jashtëm të disponueshëm. Algoritmet grupuese i ndajnë të dhënat në grupime bazuar pothuajse vetëm në ngjashmërinë e të dhënave më njëra-tjetrën, prandaj cilësohen si algoritme të drejtuar nga të dhënat [3].
Ekziston një mori algoritmesh që trajtojnë problemin e analizës së grupimeve të cilët ndryshojnë nga njëri-tjetri në aspekte të ndryshme ku ndër më kryesorët është mënyra e konceptimit të grupimit. Për shembull grupimet mund të konceptohen si bashkësi elementet e të cilëve kanë distancë të vogël nga njëri-tjetri, zona me dendësi të lartë

në hapësirën e të dhënave, intervale ose shpërndarje statistikore të veçanta. Formalisht problemi i analizës së grupimeve shprehet si një problem optimizimi multi-objektiv.
Në varësi të mënyrave të konceptimit të grupimeve, në varësi të llojeve të grupimeve që gjenerojnë, apo në varësi të karakteristikave operacionale, ekzistojë mënyra të ndryshme klasifikimi për algoritmet e grupimeve. Për shembull një nga klasifikimet kryesore është në algoritme particionalë dhe hierarkikë, ku algoritmet e grupimit hierarkikë prodhojnë një seri të ndërfutur particionesh, ndërsa algorithmet particionalë krijojnë një particion të vetëm në disa grupime [4].
Një klasifikim tjetër i rëndësishëm i algoritmeve të grupimeve i ndan ato në algoritme të grupimeve të forta dhe në algoritme të grupimeve të turbullta (gjithashtu përdoret dhe termi 'të buta'). Në algoritmet e grupimeve të forta, secila nga të dhënat ose i përket (pra është anëtare e plotë e) një grupimi ose nuk i përket (pra nuk aspak anëtare e) një grupimi. Nga ana tjetër algoritmet e grupimeve të turbullta bazohen në teorinë e logjikës së turbullt dhe bashkësive të turbullta duke krijuar particione ku një e dhënë nuk është e detyruar t'i përkasë ose plotësisht, ose aspak një grupimi, por mund të ketë anëtarësi të pjesshme (një vlerë reale midis 0 dhe 1) në disa prej grupimeve. Përqasja e turbullt në problemin e analizës së grupimeve ofron jo vetëm më shume fleksibilitet, por gjithashtu është më e natyrshme. Kështu një e dhënë që ndodhet në kufirin e dy apo më shumë grupimeve (vlera kufitare) nuk është e detyruar t'i përkasë plotësisht ndonjërit prej tyre, por njëkohësisht mund të ketë anëtarësi të pjesshme në disa prej tyre [5]. Një avantazh tjetër i algoritmeve të grupimit të turbullt është në paraqitjen e vlerave të largëta dhe të dhënave me zhurmë. Natyrisht që për këto vlera anëtarësia e plotë në ndonjë prej grupimeve nuk do të ishte modelim i mire, ndërsa në grupime të turbullta ato do të kenë anëtarësi të pjesshme në një apo disa prej grupimeve.
Përveç avantazheve nga ana kuptimore të përmendura në paragrafin e mësipërm, përqasja e turbullt ka rezultuar të ketë dhe një avantazh tjetër mjaft të rëndësishëm që është larmia e metodologjisë së zhvilluar dhe që vazhdon të zhvillohet për trajtimin e këtij problemi. Për shembull një nga algoritmet më të rëndësishëm të grupimeve të turbullta është algoritmi i turbullt i c-mesoreve i cili është formuar duke modifikuar versionin klasik (i cili bazohet në logjiken dyvlerëshe), dhe një mori algoritmesh të bazuar në logjikën e turbullt janë zhvilluar duke modifikuar parametra të ndryshëm të këtij algoritmi si metrikën e ngjashmërisë, funksionin objektiv etj.
Problemi i analizës së grupimeve është një problem mjaft kompleks. Pavarësisht se ekzistojnë disa algoritme me një sërë aplikimesh mjaft të suksesshme, është e pamundur të përcaktohet një algoritëm që të trajtojë këtë problem për të gjitha bashkësitë e të dhënave. Ekzistojnë një sërë çështjesh që duhen marrë parasysh në mënyrë të kujdesshme në lidhje me përzgjedhjen e algoritmit të grupimit si numri i grupimeve, përcaktimi i vlerave fillestare të qendrave të grupimeve (inicializimi),

metrika e ngjashmërisë (distancës), përparësia përkundrejt saktësisë apo interpretueshmërisë.
Në këtë punim synojmë të japim një studim sistematik mbi algoritmet e grupimeve të turbullta, ku t'i studiojmë teorikisht, t'i implementojmë dhe t'i krahsojmë eksperimentalisht ato. Gjithashtu të zhvillojmë linja të sintonizimint të këtyre algoritmeve si dhe teknika të avancuara hibridizuese për të përmirësuar saktësinë dhe stabilitetin e tyre. Përveç bashkësive të të dhënave standarte, këto algoritme i kemi aplikuar dhe mbi bashkësi të dhënash sintetike si dhe në bashkësi të dhënash nga bota reale.

1.2 Qëllimi

Qëllimi ynë kryesor në këtë punim është të japim një studim sistematik për trajtimin e analizës së grupimeve bazuar në logjikën e turbullt, të shoqëruar me implementime dhe krahasime eksperimentale mbi një sërë bashkësish të dhënash. Në funksion të qëllimit tonë kryesor vemë në dukje këto objektiva:

- Hulumtimi i hollësishëm i algoritmeve ekzistues për trajtimin e analizës së grupimeve të turbullt.
- Hulumtim i hollësishëm i teknikave sintonizuese mbi këto algoritme për përmirësimin e saktësisë dhe stabilitetit.
- Studim i teknikave të përzgjedhjes së veçorive të nevojshjme në trajtimin e problemit të analizës së grupimeve në bashkësi të dhënash shumë-përmasore.
- Studim i indekseve vlerësues të cilësisë së grupimeve (i njohur si problemi i verifikimit të grupimeve)
- Studim i teknikave të përshtatjes së algoritmeve të analizës së turbullt të grupimeve për të vepruar në kushte të mësimit gjysmë të mbikëqyrur.
- Studim i teknikave të përdorimit të metodave ansambël (homogjene dhe heterogjene) në grupime për të ndërtuar modele grupues me saktësi dhe stabilitet më të lartë
- Implementime të teknikave të lartpërmendura në bashkësi të dhënash standarte, në bashkësi të dhënash sintetike dhe në bashkësi të dhënash nga bota reale.

1.3 Vështrim i përgjithshëm i përmbajtjes

Pasi japim kornizën e përgjithshme teorike, motivimin dhe qëllimet kryesore të punimit gjatë këtij kapitulli (kapitulli i parë), teza vazhdon më tej me kapitullin e dytë në të cilin trajtohen themelet teorike të logjikës së turbullt, si bashkësitë e turbullta,

funksionet e anëtarësisë, relacionet e turbullta, operatorët, procedurat e shturbullimit si dhe zgjatime të logjikës së turbullt si logjika e turbullt intuitive dhe logjika e turbullt e llojit të dytë.

Në kapitullin e tretë do të paraqesim një studim të hollësishëm rreth fushës së analizës së grupimeve si një prej komponentëve më të rëndësishme të mësimit pas mbikqyrje. Do të diskutohen algoritme që e trajtojnë problemin e analizës së grupimeve në disa forma si algortimet klasike dhe algoritmet e turbullt (ndryshe quhen dhe algoritmet e "fortë" dhe algoritmet e "butë"), algoritmet particionale dhe algoritmet hierakike. Boshti kryesor i studimit do të jetë algoritmi i grupimit të turbullt të mesoreve dhe një sërë variantesh të tij si algoritmet Gustafson-Kessel, Gath-Geva, grupimi i turbullt i mesoreve bazuar në bërthama etj. Gjithashtu do të jepen variante për përzgjedhjen e komponentëve kryesorë të këtyre algoritmeve si për shembull metrika e distancës, koeficentët e turbullsisë, funksionet bërthamë etj.

Në kapitullin e katërt do të trajtojmë problemin e vlerësimit të cilësisë së grupimeve të përftuara (i njohur si problemi i verifikimit të grupimeve) si dhe do të zhvillojmë teknika të ndryshme sintonizuese për algoritmet e grupimit të turbullt të shoqëruara me një sërë procedurash eksperimentale vlerësuese dhe krahasuese të zbatuara mbi bashkësi të dhënash nga bota reale dhe sintetike.

Në kapitullin e pestë do të paraqitet një studim teorik i teknikave të përshtatjes së algoritmeve të analizës së turbullt të grupimeve për të vepruar në kushte të mësimit gjysmë-të mbikqyrur si dhe një studim eksperimental i efiçiencës dhe shkallës së saktësisë së algoritmeve të grupimit të turbullt në kushte të mësimit gjysmë-të mbikqyrur.

Në kapitullin e gjashtë diskutohet mbi metodat ansambël (homogjene dhe heterogjene) në grupime. Këto metoda synojnë së pari ndërtimin e modeleve grupues me saktësi dhe stabilitet të lartë, si dhe së dyti të shmangin disa probleme kritike si si ndjeshmëria ndaj përzgjedhjes së vlerave fillestare për qendrat e grupimeve, ndjeshmëria ndaj vlerave të largëta dhe ndjeshmëria ndaj zhurmës. Në këtë kapitull pasi të trajtojmë parimet e përgjithshme të projektimit të një metode ansambël, do të propozojmë disa metoda ansambël si dhe do t'i zbatojmë eksperimentalisht ku të vlerësojmë dhe krahasojmë shkallës së saktësisë dhe stabilitetit së tyre.

Në kapitullin e shtatë, i cili është kapitulli përmbyllës, do të nxirren konkluzionet dhe vihen dukje kontributet e këtij punimi. Gjithashtu sugjerohen mundësi të reja të kërkimit shkencor si vazhdë e këtij punimi.

2. LOGJIKA E TURBULLT

Koncepti i pasigurisë është mjaft i rëndësishëm në shumë fusha të shkencës. Arsyeja kryesore është pamundësia për të modeluar dhe trajtuar skenare të botës reale brenda kornizave të modeleve rigorozë. Deri në fundin e shekullit XIX pasiguria konsiderohej e padëshirueshme në proceset e kërkimit shkencor dhe gjithnjë synohej shmangia e pothuajse të gjithë formave të saj si mungesa e precizionit, mungesa e specifikave, rastësia etj. Ndërsa me trajtimin e problemeve komplekse në fusha si fizika në nivelet molekulare, statistika, makroekonomia, teoria e informacionit, meteorologjia etj, gjithnjë e më shumë u ndje nevojshmëria për mekanizma që veprojnë në kushte të pasigurisë.

Logjika e turbullt është një fushë relativisht e re studimi, e cila mundëson trajtimin e disa formave të pasigurisë. Zanafilla e logjikës së turbullt daton në vitin 1965 me prezantimin e teorisë së bashkësive të turbullta nga Lotfi Zadeh (i cili tashmë është i njohur si "babai" i logjikës së turbullt) [6]. Ndërkohë që në logjikën klasike çdo pohim është në mënyrë precize ose i vërtetë (vlera 1) ose i rremë (vlera 0), një pohim në logjikën e turbullt mund të ketë vërtetësi të pjesshme (pra një vlerë reale midis 0 dhe 1).

Si shembull ilustrues le të marrim në konsideratë dy gota uji, ku gota e parë eshtë 80% e mbushur me ujë, ndërsa gota e dytë është 25% e mbushur me ujë. Atëherë në logjikën e turbullt pohimi "Gota është plot" për gotën e parë do të kishte shkallë vërtetësie 0.8, ndërsa për gotën e dytë i njëjti pohim do të kishte shkallë vërtetësie 0.25. Gjithashtu pohimi i anasjelltë "Gota është bosh" për gotën e parë do të kishte shkallë vërtetësie 0.2, ndërsa për gotën e dytë 0.75. Nga ana tjetër është e qartë se në logjikën klasike për të dyja rastet pohimi "Gota është bosh" do të ishte i rremë, pra do të kishte shkallë vërtetësie 0. Ndonëse shembulli ishte mjaft i thjeshtë, ai është i mjaftueshëm për të vënë në dukje faktin se shprehja e shkallës së vërtetësisë në logjikën e turbullt ishte më informuese se shprehja e shkallës së vërtetësisë në logjikën klasike. Konkretisht në logjikën klasike nuk pati dallim për shkallën e vërtetësisë së pohimit midis dy gotave në fjalë, ndërsa në logjikën e turbullt dallimi ndërmjet dy rasteve ishte i dukshëm.

Përgjithësisht logjika e turbullt ofron mjaft fleksibilitet në mënyrën si perceptohet, modelohet dhe përpunohet informacioni duke bërë që ajo të jetë një mjet gjerësisht i përdorur për trajtimin e problemeve në kushte të prezencës së pasigurisë dhe ka gjetur mjaft zbatime të suksesshme në një sërë fushash si teori kontrolli, bioinformatikë, mjekësi, përpunim imazhesh, ekonomi, shkenca konjitive etj.

2.1 Bashkësitë e turbullta

Siç dihet, në bashkësitë klasike për çdo element të bashkësisë universale X, të qenit apo jo element i një bashkësie të caktuar klasike është i prerë: ose është plotësisht element i saj, ose nuk është aspak element i saj. Nga ana tjetër, bashkësitë e turbullta janë një përgjithësim i bashkësive klasike, tek të cilat elementet e bashkësisë universale X nuk është e thënë të jenë plotësisht ose aspak anëtarë të saj, por mund të kenë një vlerë anëtarësie të pjesshme (një vlerë reale midis 0 dhe 1).

Për t'u shprehur në terma më formalë, prezantohet funksioni i anëtarësisë i formës $\mu_A: X \to [0,1]$ i cili karakterizon bashkësinë e turbullt A duke pasqyruar çdo element të bashkësisë universale X në një vlerë të segementit real [0,1]. Natyrisht që dhe bashkësitë klasike (si raste të veçanta të bashkësive të turbullta) karakterizohen nga funksione anëtarësie, vetëm se në këtë rast çdo element i bashkësisë universal X pasqyrohet në një nga dy vlerat e bashkësisë diskrete $\{0,1\}$.

Konvencioni për shënimin e elementeve të një bashkësie të turbullt, kur ajo është diskrete dhe e fundme është:

$$A = \left\{\frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + \cdots + \frac{\mu_A(x_n)}{x_n}\right\} = \sum_i \frac{\mu_A(x_i)}{x_i}$$

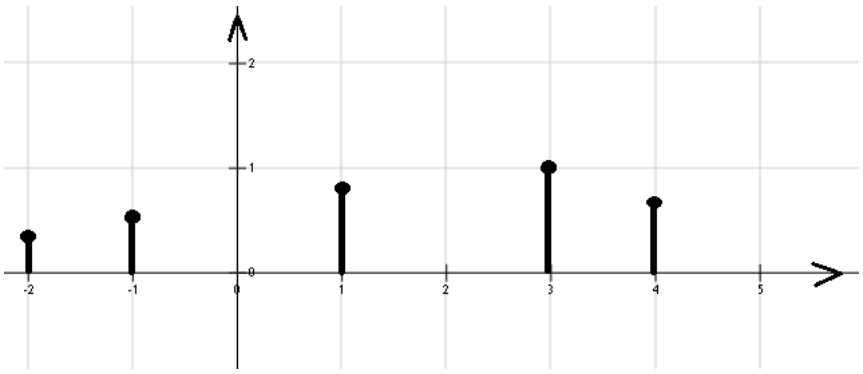
Ndërsa kur bashkësia e turbullt është e vazhdueshme, konvencioni i shënimit është:

$$A = \left\{\int \frac{\mu_A(x)}{x}\right\}$$

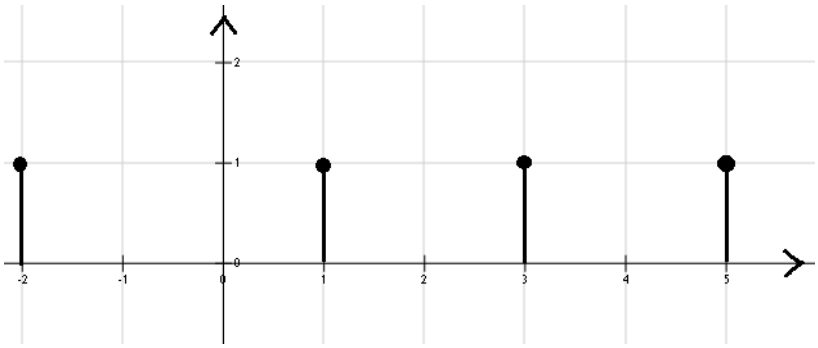
Në të dyja këto shënime është për t'u vënë në dukje se vija horizontale nuk është në kuptimin e thyesës, por thjesht një vijë ndarëse, kurse simboli '+' nuk kryen shumën algjebrike por është një operator agregues (që shërben për grupimin e të dhënave) [7].

Le të japim një shembull bashkësie te turbullt diskrete $A_1 = \left\{\frac{0.4}{-2} + \frac{0.6}{-1} + \frac{0.8}{1} + \frac{1}{3} + \frac{0.7}{4}\right\}$. Madje edhe një bashkësi klasike si $A_2 = \{-2, 1,3,5\}$ mund të paraqitet në shënimin për bashkësitë e turbullta si $A_2 = \left\{\frac{1}{-2} + \frac{1}{1} + \frac{1}{3} + \frac{1}{5}\right\}$ (të gjithë elementet kanë anëtarësi të plotë, pra me vlerë 1).

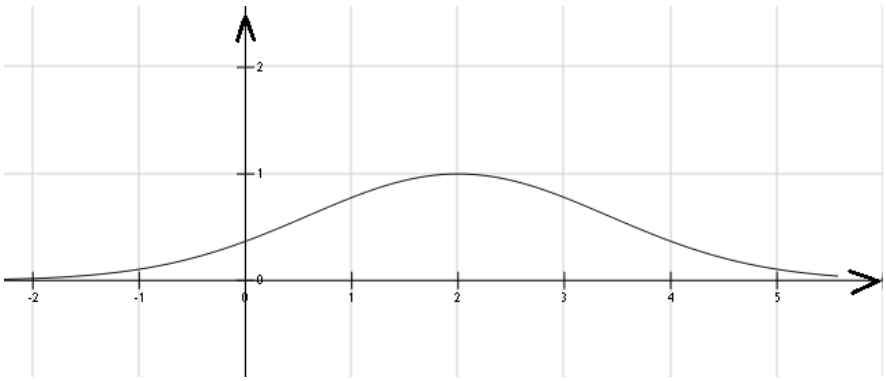
Gjithatu le të japim për qëllime ilustrimi një shembull bashkësie të turbullt të vazhdueshme $A_3 = \left\{\int \frac{e^{-\frac{1}{4}(x-2)^2}}{x}\right\}$ si dhe një shembull bashkësi klasike të vazhdueshme $A_4 = [-\infty, 1] \cup [2,5]$. Për të katërta këto bashkësi po japim grafikët e funksioneve të anëtarësisë ne vijim, ku dallohet qartë se për bashkësitë klasike vlerat e funksionit mund të jenë ose 0, ose 1, ndërsa për bashkësitë e turbullta vlerat e funksionit variojnë brenda segmentit [0,1].



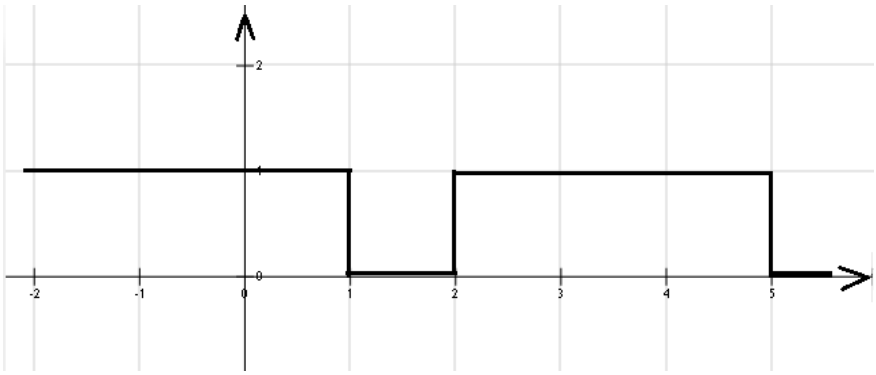
Figurë 1 - Funksioni i anëtarësisë për bashkësinë e turbullt diskrete A_1



Figurë 2 - Funks ioni i anëtarësisë për bashkësinë klasike diskrete A_2

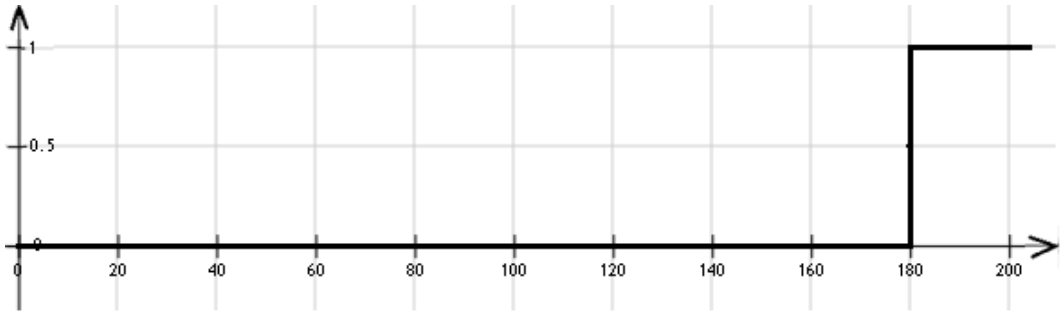


Figurë 3 - Funksioni i anëtarësisë për bashkësinë e turbullt të vazhdueshme A_3



Figurë 4 - Funksioni i anëtarësisë për bashkësinë klasike të vazhdueshme A_4

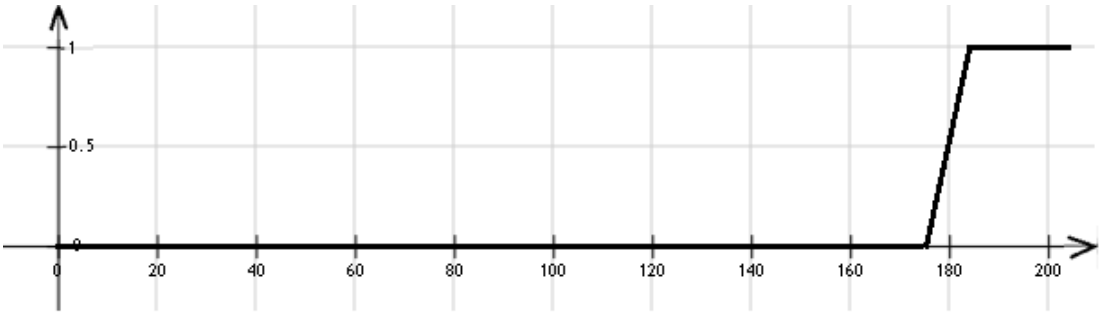
Bashkësitë e turbullta kanë një aftësi shprehëse mjaft të mirë veçanërisht kur bëhet fjalë për të paraqitur bashkësi të botës reale. Për shembull supozojmë se duam të shprehim bashkësinë e njerëzve të gjatë. Në rast se do të vepronim me anë të bashkësive klasike, do të na duhej të vendosim një kufi të prerë i cili do të përcaktojë nëse një element i përket bashkësisë në fjalë apo jo. Në shembullin tonë vlerën e këtij kufiri po e vendosim 180 (cm). Atëherë funksioni i anëtarësisë i bashkësisë sonë do të jetë:



Figurë 5 - Bashkësia e njerëzve të gjatë si bashkësi klasike

Pra në paraqitjen si bashkësi klasike personat me gjatësi nën 180 cm nuk janë aspak anëtarë të bashkësisë së njerëzve të gjatë (shkalla e anëtarësisë së tyre është 0), ndësa personat me gjatësi të paktën 180 cm konsiderohen anëtarë të plotë (shkalla e anëtarësisë së tyre është 1). Paraqitja e kësaj bashkësie të botës reale në këtë formë nuk ofron fleksibilitet, konkretisht një person me gjatësi 179 cm nuk i përket bashkësisë së njerëzve të gjatë, ndërsa një person me gjatësi 180 cm i përket kësaj bashkësie, ndërkohë që ndryshimi midis dy personave në fjalë është mjaft i vogël. (Vemë në dukje që kjo mungesë fleksibiliteti nuk ka të bëjë me vlerën e përzgjedhur të kufirit. Sikur vlera e kufirit të ishte 185 cm, atëherë do të kishim që një person 184 cm nuk i përket bashkësisë së njerëzve të gjatë ndërkohë që një person me gjatësi 185 cm i përket bashkësisë së njerëzve të gjatë).

Duke përdorur bashkësitë e turbullta, mund ta paraqesim bashkësinë e njerëzve të gjatë në një mënyrë më të natyrshme dhe më fleksibël me një funksion anëtarësie si në figurën vijuese:



Figurë 6 - Bashkësia e njerëzve të gjatë si bashkësi e turbullt

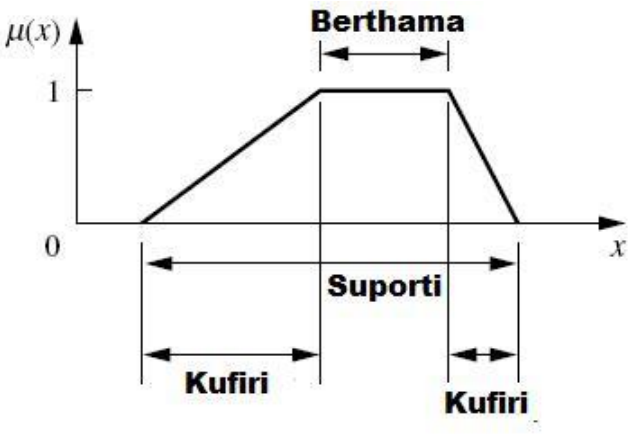
Pra personat me gjatësi nën 175 cm nuk janë aspak anëtarë (shkalla e anëtarësisë është 0) të bashkësisë së njerëzve të gjatë, personat me gjatësi mbi 185 cm janë anëtarë të plotë (shkalla e anëtarësisë është 1) të bashkësisë së njerëzve të gjatë, ndërsa personat me gjatësi midis 175 dhe 185 cm kanë anëtarësi të pjesshme. Në këtë rast shmanget dhe problemi i përmendur më parë ku për persona me gjatësi pothuajse të njejtë mund të ketë ndryshime drastike. Konkretisht një person me gjatësi 179 cm ka anëtarësi 0.4 në bashkësinë e njerëzve të gjatë ndërsa një person me gjatësi 180 cm ka anëtarësi 0.5 në këtë bashkësi (pra dhe ndryshimi në anëtarësi është i vogël dhe jo drastik).
Në mënyrë të ngjashme bashkësitë e turbullta mund të përdoren me sukses për të shprehur bashkësi të tjera të botës reale për të cilat një ndarje e prerë nuk do të ishte e natyrshme si për shembull bashkësia e njerëzve të rinj (ose bashkësia e njerëzve të vjetër), bashkësia e personave të pasur (ose të varfër), bashkësia makinave të shtrenjta (ose të lira), bashkësia e vendeve apo ditëve të nxehta (ose të ftohta), bashkësia e ditëve me mot të kthjellët (ose te vrenjtur) etj.

2.1.1 Funksionet e anëtarësisë

Në këtë seksion do të paraqesim disa nga karakteristikat kryesore të funsioneve të anëtarësisë si dhe do të përshkruajmë disa nga funksionet e anëtarësisë që përdoren më gjerësisht. Siç përmendëm në seksionin e mëparshëm, funksioni i anëtarësisë $\mu_A: X \rightarrow [0,1]$ i një bashkësie të turbullt A, pasqyron çdo element të bashkësisë universale në një vlerë reale midis 0 dhe 1. Tri nga karakteristikat kryesore të funksionit të anëtarësisë janë bërthama, suporti dhe kufiri, për të cilat le të japim përkufizimet përkatëse [8]:

- Bërthama e funksionit të anëtarësisë për një bashkësi të turbullt A është bashkësia e vlerave të cilat kanë anëtarësi të plotë në këtë bashkësi, pra bashkësia e të gjithë vlerave x për të cilat $\mu_A(x) = 1$. Simbolikisht e shprehim si $berthama(A) = \{x|\mu_A(x) = 1\}$.
- Suporti i funksionit të anëtarësisë për një bashkësi të turbullt A është bashkësia e vlerave të cilat kanë anëtarësi pozitive, pra bashkësia e të gjithë vlerave x, për të cilat $\mu_A(x) > 0$. Simbolikisht e shprehim si $suporti(A) = \{x|\mu_A(x) > 0\}$.
- Kufiri i funksionit të anëtarësisë është bashkësia e vlerave që kanë anëtarësi të ndryshme nga 0 dhe të ndryshme nga 1, pra bashkësia e të gjithë vlerave x për të cilat $0 < \mu_A(x) < 1$. Këto janë elementët që kanë anëtarësi te pjesshme në bashkësinë A. Simbolikisht e shprehim si $kufiri(A) = \{x|0 < \mu_A(x) < 1\}$.

Në figurën në vijim i ilustrojmë këto koncepte:



Figurë 7 - Karakteristikat e funksionit të anëtarësisë

Përcaktimi i funksionit të anëtarësisë është një proces delikat për realizimin e të cilit janë zhvilluar një sërë teknikash të automatizuara, por gjithashtu në raste specifike preferohet të projektohet në bazë të ekspertizës njerëzore në fushën rreth së cilës do të implementohet sistemi i bazuar në logjikën e turbullt. Faktorë si efiçienca, thjeshtësia, dhe përshtatshmëria gjithnjë merren në konsideratë gjatë përzgjedhjes së funksionit të anëtarësisë. Analiza e turbullt e grupimeve të cilën do ta studiojmë me kujdes në kapitujt në vijim është një mjet i sofistikuar për eksplorimin e të dhënave, por në një këndvështrim të ngushtë mund të shihet si mjet përcaktues i funksionit të anëtarësisë për grupimet. Përgjithësisht përcaktimi i funksionit të anëtarësisë nuk është unik, pra për një skenar të caktuar mund të ketë disa funksione anëtarësie të cilat operojnë me sukses. Në vijim po përshkruajmë disa nga funksionet e anëtarësisë të përdorur më gjerësisht [9]:

- Funksioni trekëndësh: $$\mu_A(x) = \begin{cases} 0 & per\ x < a \\ \frac{x-a}{b-a} & per\ a \le x < b \\ \frac{c-x}{c-b} & per\ b \le x < c \\ 0 & per\ x \ge c \end{cases}$$

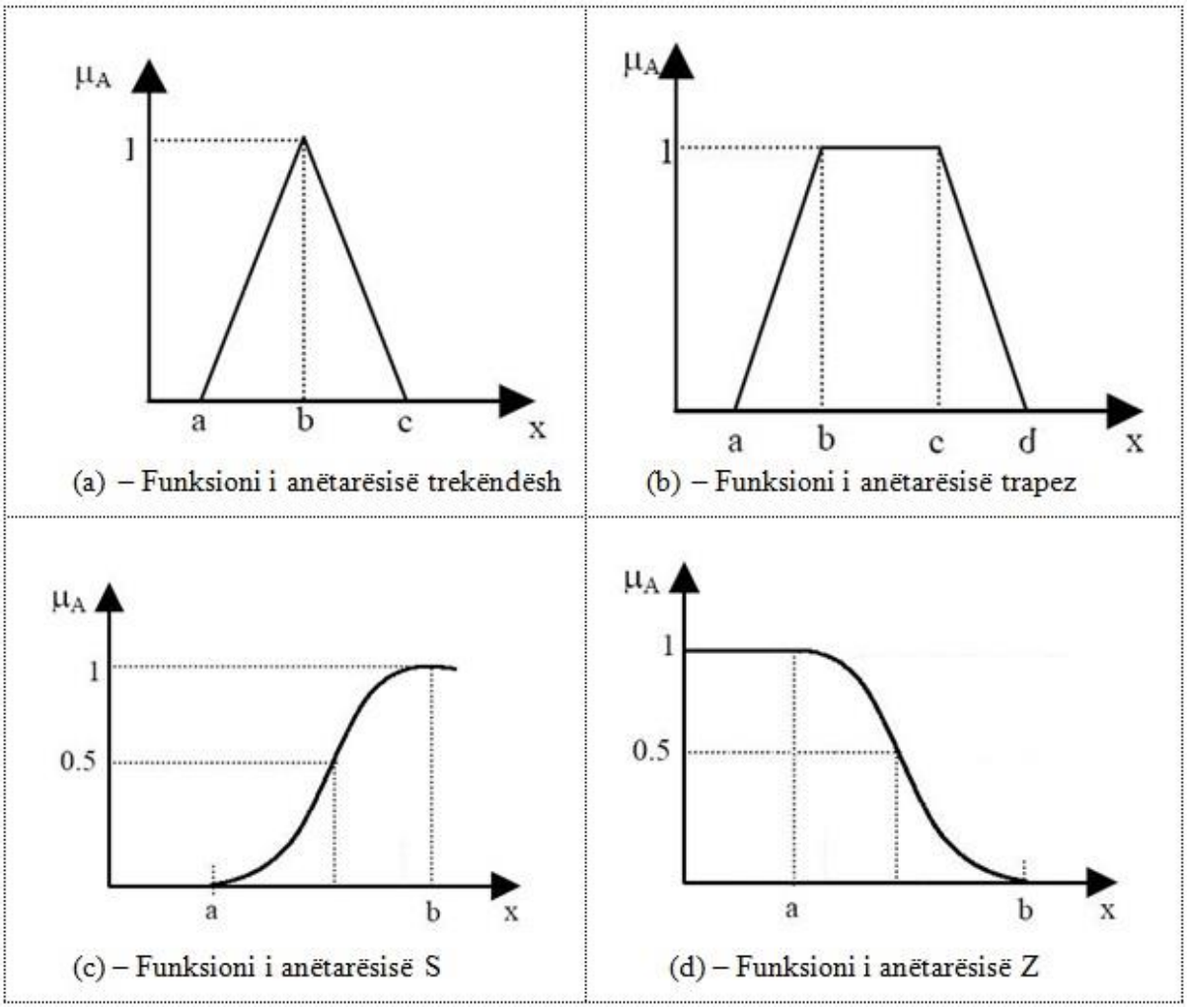
- Funksioni trapez: $$\mu_A(x) = \begin{cases} 0 & per\ x < a \\ \frac{x-a}{b-a} & per\ a \le x < b \\ 1 & per\ b \le x < c \\ \frac{d-x}{d-c} & per\ c \le x < d \\ 0 & per\ x \ge d \end{cases}$$

- Funksioni S : $\mu_A(x) = \begin{cases} 0 & per\ x < a \\ 2\left(\frac{x-a}{b-a}\right)^2 & per\ a \leq x < (a+b)/2 \\ 1 - 2\left(\frac{b-x}{b-a}\right)^2 & per\ (a+b)/2 \leq x < b \\ 1 & per\ x \geq b \end{cases}$
- Funksioni Z: $\mu_A(x) = \begin{cases} 1 & per\ x < a \\ 1 - 2\left(\frac{x-a}{b-a}\right)^2 & per\ a \leq x < (a+b)/2 \\ 2\left(\frac{b-x}{b-a}\right)^2 & per\ (a+b)/2 \leq x < b \\ 1 & per\ x \geq b \end{cases}$
- Funksioni Gausian: $\mu_A(x) = e^{-\frac{(x-b)^2}{2a^2}}$
  Bashkësia A_3 që kemi dhënë më parë në figurën 3 është një rast bashkësie e turbullt me funksion anëtarësie Gausian me parametra $a = 1.41$ dhe $b = 2$.
- Funksioni këmbanë: $\mu_A(x) = \frac{1}{1+\left|\frac{x-c}{a}\right|^{2b}}$ ku $a, b, c \in R$ dhe $b > 0$.
- Funksioni sigmoid: $\mu_A(x) = \frac{1}{1+e^{-a(x-b)}}$ ku $a, b \in R$ Në varësi të shenjës së parametrit a, funksionin sigmoid mund të jetë i hapur në të majtë ose në të djathtë, kështu që është një funksion i përshtatshëm për paraqitjen e bashkësive të botës reale si: “shumë i madh”, “shumë i ftohtë” etj.

Në bazë të funksioneve të anëtarësisë, konceptet themelore në lidhje me bashkësitë klasike si bashkësia boshe, barazia, nënbashkësitë, kardinaliteti etj mund t’i përkufizojmë edhe për bashkësitë e turbullta:

- Një bashkësi e turbullt A quhet boshe nëse $\mu_A(x) = 0, \forall x \in X$
- Dy bashkësi të turbullta A dhe B janë të barabarta nëse $\mu_A(x) = \mu_B(x)$, $\forall x \in X$
- Një bashkësi e turbullt A është nënbashkësi e një bashkësie të turbullt B nëse $\mu_A(x) \leq \mu_B(x), \forall x \in X$
- Një bashkësi e turbullt A quhet normale nëse $\exists x \in X$ i tillë që $\mu_A(x) = 1$ (pra nëse bërthama e tij nuk është bashkësi boshe).
- Kardinaliteti (numri i elementëve) për një bashkësi të turbullt shprehet si shuma e vlerave të anëtarësisë së elementëve të saj. Pra, për $A = \left\{\frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + \cdots + \frac{\mu_A(x_n)}{x_n}\right\}$, kemi $card_A = \mu_A(x_1) + \mu_A(x_2) + \cdots + \mu_A(x_n)$

Në figurën vijuese japim paraqitjen grafike të formave kryesore të funksioneve të anëtarësië të përmendura më lart:



Figurë 8 - Disa nga format kryesore të funksioneve të anëtarësisë

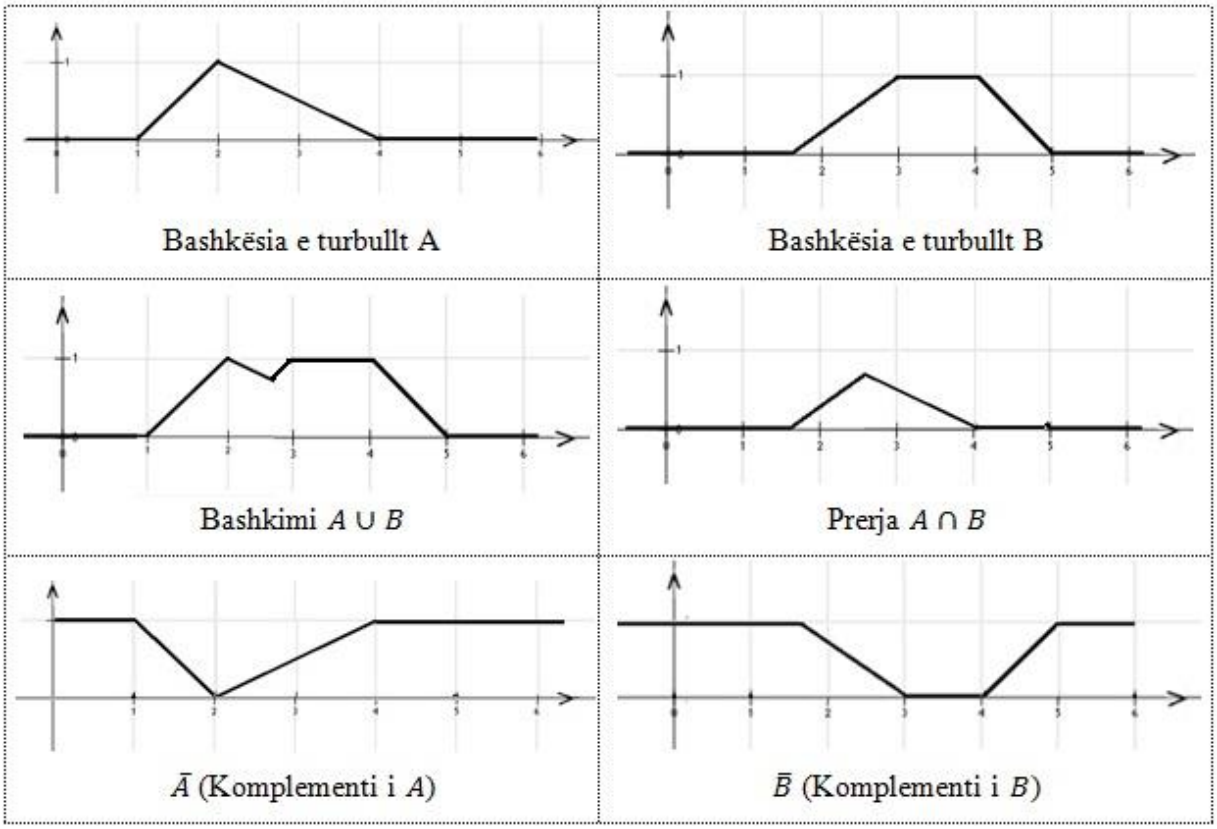
2.1.2 Veprimet mbi bashkësitë e turbullta dhe vetitë e tyre

Në bashkësitë e turbullta, ashtu si në bashkësitë klasike, tri veprimet themelore që zbatohen mbi to janë bashkimi, prerja dhe komplementi. Supozojmë se A dhe B janë dy bashkësi të turbullta me funksione anëtarësie përkatësisht $\mu_A(x)$ dhe $\mu_B(x)$. Atëherë tri veprimet themelore i përkufizojmë si më poshtë [7]:

- Bashkimi i bashkësive të turbullta A dhe B është bashkësia e turbullt $A \cup B$, me funksion anëtarësie $\mu_{A\cup B}(x) = \max(\mu_A(x),\ \mu_B(x))$
- Prerja e bashkësive të turbullta A dhe B është bashkësia e turbullt $A \cap B$, me funksion anëtarësie $\mu_{A\cap B}(x) = \min\big(\mu_A(x),\ \mu_B(x)\big)$

- Komplementi i bashkësisë së turbullt A është bashkësia e turbullt $\bar{A}$, me funksion anëtarësie $\mu_{\bar{A}}(x) = 1 - \mu_A(x)$

Në figurën vijuese i ilustrojmë tri veprimet themelore mbi bashkësitë e turbullta:



Figurë 9 - Veprimet themelore mbi bashkësitë e turbullta

Vemë në dukje se këto tri veprime funksionojnë njësoj edhe mbi bashkësitë klasike në të cilat vlerat e anëtarësisë së elementëve i përkasin bashkësisë diskrete {0,1}, pra bashkimi, prerja dhe komplementi në bashkësitë e turbullta janë përgjithësime të bashkimit, prerjes dhe komplementit në bashkësitë klasike. Këto veprime në bashkësitë e turbullta gjithashtu gëzojnë shumicën e vetive të bashkësive klasike si vetia e ndërrimit, shoqërimit, përdasimit, njëshmërisë, dominimit, identitetit, komplementit të dyfishtë dhe De Morganit [7]. Më konkretisht këto veti janë të përmbledhura në tabelën në faqen vijuese.

Në krahasim me bashkësitë klasike, e vetmja veti e cila nuk plotësohet nga bashkësitë e turbullta është vetia e komplementit, pra në përgjithësi për një bashkësi të turbullt A: $A \cup \bar{A} \neq X$ (bashkimi i një bashkësie të turbullt me komplementin e saj nuk jep bashkësinë universale) dhe $A \cap \bar{A} \neq \Phi$ (prerja e një bashkësie të turbullt me komplementin e saj nuk jep bashkësinë boshe).

Identiteti	Emri
$A \cup B \equiv B \cup A$ $A \cap B \equiv B \cap A$	Vetia e ndërrimit
$A \cup (B \cup C) \equiv (A \cup B) \cup C$ $A \cap (B \cap C) \equiv (A \cap B) \cap C$	Vetia e shoqërimit
$A \cup (B \cap C) \equiv (A \cup B) \cap (A \cup C)$ $A \cap (B \cup C) \equiv (A \cap B) \cup (A \cap C)$	Vetia e përdasimit
$A \cup A \equiv A$ $A \cap A \equiv A$	Vetia e njëshmërisë
$A \cup X \equiv X$ $A \cap \Phi \equiv \Phi$	Vetia e dominimit
$A \cup \Phi \equiv A$ $A \cap \mathrm{X} \equiv \mathrm{A}$	Vetia e identitetit
$\bar{\bar{A}} \equiv A$	Vetia e komplementit të dyfishtë
$\overline{A \cup B} \equiv \bar{A} \cap \bar{B}$ $\overline{A \cap B} \equiv \bar{A} \cup \bar{B}$	Vetia e De Morganit

Tabelë 1 - Vetitë e veprimeve në bashkësitë e turbullta

Veprimet mbi bashkësitë e turbullta që paraqitëm, cilësohen si veprimet standarde dhe përgjithësojnë veprimet përkatëse në bashkësitë klasike, por ato nuk janë e vetmja mënyrë si mund të përgjithësohen veprimet në fjalë. Ekziston një klasë e gjerë funksionesh të cilët mund të konsiderohen si përgjithësime të turbullta të veprimeve standarde në bashkësitë klasike. Funksionet të cilët plotësojnë kriteret për t'u konsideruar si prerje të turbullta në literaturë njihen si norma trekëndore (t-norma), ndërsa funsionet të cilët plotësojnë kriteret për t'u konsideruar si bashkime të turbullta njihen si bashkënorma trekëndore (t-bashkënorma). Në spektrin e gjerë të t-normave, prerja standarde zë një vend të veçantë pasi ajo përbën prerjen më të madhe të mundshme që mund të prodhohet për dy bashkësi të turbullta të çfarëdoshme. Gjithashtu në spektrin e gjerë të t-bashkënormave, bashkimi standard zë një vend të veçantë pasi përbën bashkimin më të vogël të mundshëm që mund të prodhohet për dy bashkësi të turbullta.

Një t-normë përkufizohet si një funksion $T(x, y)$ i cili pasqyron vlerat nga $[0,1]^2$ në $[0,1]$ dhe plotëson kriteret [10]:

- $T(0,0) = T(0,1) = T(1,0) = 0$ (vetia e vlerave kufitare)
- $T(x, 1) = x, \ \forall x \in [0,1]$ (ekzistenca e elementit neutral)
- $T(x, y) = T(y, x), \ \forall x \forall y \in [0,1]$ (vetia e ndërrimit)
- $T(x, T(y, z) = T(T(x, y), z), \ \forall x \forall y \forall z \in [0,1]$ (vetia e shoqërimit)

- $T(u,v) \leq T(x,y)$, për çdo $0 \leq u \leq x \leq 1$ dhe $0 \leq v \leq y \leq 1$ (vetia e monotonisë)

Disa shembuj t-normash të cilat janë studiuar hollësisht dhe janë zbatuar me sukses në praktikë janë [10], [11]:

- t-norma minimale (e cila e ka origjinën në logjikën e Gödel-Dummet) përkufizohet si $T(x,y) = min(x,y)$
- t-norma e Lukasiewicz (e cila e ka origjinën në logjikën e Lukasiewicz) përkufizohet si $T(x,y) = max(x+y-1,0)$
- t-norma prodhim e cila përkufizohet si $T(x,y) = x \cdot y$
- t-norma drastike e cila përkufizohet si $T(x,y) = min(x,y)$ nëse $max(x,y) = 1$, përndryshe $T(x,y) = 0$. Kjo t-normë përfaqëson prerjen më të vogël të mundshme për dy bashkësi të turbullta të çfarëdoshme.
- t-norma e minimumit nilpotent e cila përkufizohet si $T(x,y) = min(x,y)$ nëse $x + y > 1$, përndryshe $T(x,y) = 0$.
- t-norma Hamacher e cila përkufizohet si $T(0,0) = 0$ dhe $T(x,y) = xy/(x + y - xy)$ kur të paktën një nga variablat x, y është i ndryshëm nga 0. Kjo t-normë mund të përgjithësohet gjithashtu si $T(x,y) = xy/(r + (1-r)(x + y - xy))$ për çdo $r > 0$.

Një t-bashkënormë përkufizohet si një funksion $S(x,y)$ i cili pasqyron vlerat nga $[0,1]^2$ në [0,1] dhe plotëson kriteret [12]:

- $S(x,0) = x$, $\forall x \in [0,1]$ (ekzistenca e elementit neutral)
- $S(x,y) = S(y,x)$, $\forall x \forall y \in [0,1]$ (vetia e ndërrimit)
- $S(x,S(y,z) = S(S(x,y),z)$, $\forall x \forall y \forall z \in [0,1]$ (vetia e shoqërimit)
- $S(u,v) \leq S(x,y)$, për çdo $0 \leq u \leq x \leq 1$ dhe $0 \leq v \leq y \leq 1$ (vetia e monotonisë)

Disa shembuj t-bashkënormash të cilat janë studiuar hollësisht dhe janë zbatuar me sukses në praktikë janë [13]:

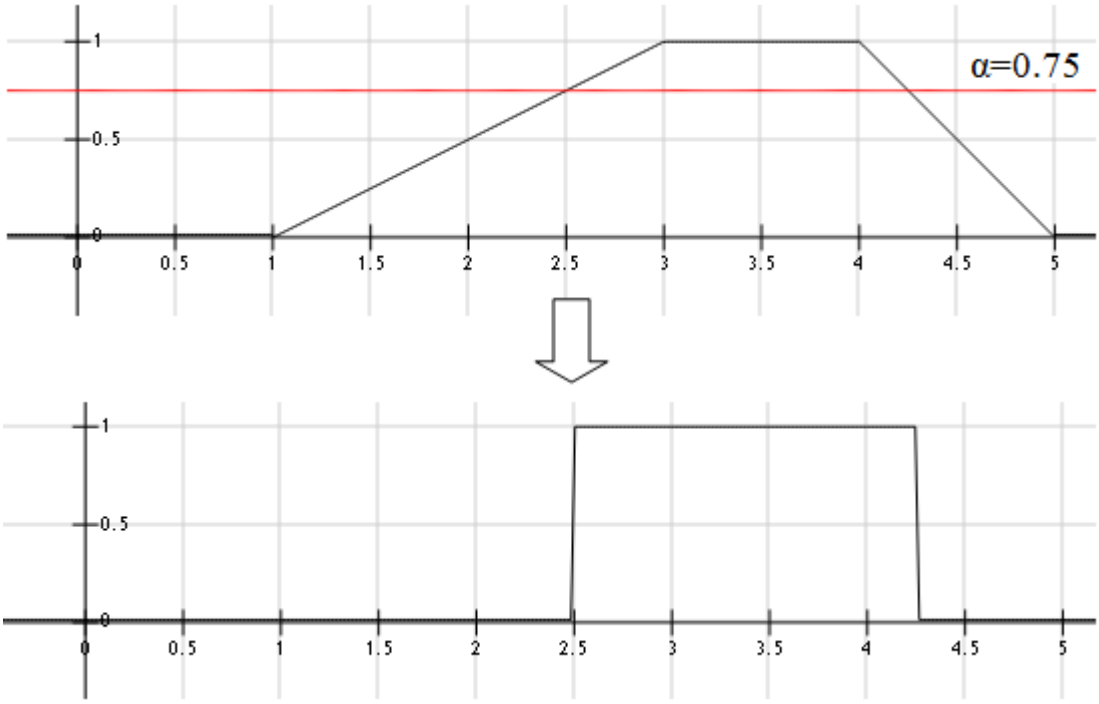
- t-bashkënorma maksimale (e cila e ka origjinën në logjikën e Gödel-Dummet) dhe përkufizohet si $S(x,y) = max(x,y)$
- t- bashkënorma e Lukasiewicz (e cila e ka origjinën në logjikën e Lukasiewicz) dhe përkufizohet si $S(x,y) = min(x+y,1)$
- t-bashkënorma prodhim (ose probabilitare) përkufizohet si $S(x,y) = x + y - x \cdot y$
- t-bashkënorma drastike përkufizohet si $S(x,y) = max(x,y)$ nëse $min(x,y) = 0$, përndryshe $S(x,y) = 1$. Kjo t-bashkënormë përfaqëson bashkimin më të madh të mundshëm për dy bashkësi të turbullta të çfarëdoshme.

- t-norma e maksimumit nilpotent e cila përkufizohet si $T(x,y) = max(x,y)$ nëse $x + y < 1$, përndryshe $T(x,y) = 1$.

Përgjithësisht për çdo t-normë $T(x,y)$ mund të gjenerojmë një t-bashkënormë $S(x,y)$ me anë të relacionit $S(x,y) = 1 - T(1-x, 1-y)$.

2.1.3 Transformimi në bashkësi klasike

Pavarësisht se bashkësitë e turbullta ofrojnë me tepër fleksibilitet, në raste jo të rralla mund të ndodhemi në rrethana ku bashkësia rezultat të kërkohet të jetë e shprehur si bashkësi klasike. Për të realizuar këtë, ekzistojnë disa teknika transformimi nga bashkësi të turbullta në bashkësi klasike (të njohura dhe si teknikat e shturbullimit në bashkësi klasike). Në këtë seksion do të paraqesim një nga teknikat e cila është njëkohësisht gjerësisht e përdorur dhe lehtësisht e implementueshme: teknika e prerjeve α (alfa). Kjo teknikë funksionon duke përzgjedhur një vlerë prag (vlera α) në intervalin [0,1] dhe bashkësia klasike rezultat do të përmbajë të gjithë elementët e bashkësisë së turbullt që kanë vlerë anëtarësie më të madhe ose të barabartë me α [14]. Bashkësia klasike e formuar emërtohet A_α, pra $A_\alpha = \{x|\mu_A(x) \geq \alpha\}$. Në figurën e mëposhtme japim një bashkësi të turbullt A dhe bashkësinë klasike $A_{0.75}$të formuar duke zbatuar prerjen α (për $\alpha = 0.75$) mbi bashkësinë në fjalë.



Figurë 10 - Transformimi në bashkësi klasike me anë të prerjes α

Nga përkufizimi vemë në dukje se për rastin kur $\alpha = 0$ kemi $A_0 = X$ sepse çdo element i universit ka një anëtarësi më të madhe ose të barabartë me 0 në bashkësinë e turbullt A. Gjithashtu mund të përdoret dhe vlera $\alpha = 0^+$ (e ngjashme me ε e përdorur gjerësisht në fushën e analizës matematikore), e cila përkufizohet si një vlerë shumë pak më e madhe se 0, pra bashkësia A_{0^+} përmban ato elemente që kanë anëtarësi më të madhe se 0 në bashkësinë e turbullt A (suporti i kësaj bashkësie). Në vijim po japim disa nga vetitë kryesore të prerjeve α:

- $(A \cup B)_\alpha = A_\alpha \cup B_\alpha$
- $(A \cap B)_\alpha = A_\alpha \cap B_\alpha$
- për çdo $0 \leq \alpha_1 < \alpha_2 \leq 1$, është i vërtetë $A_{\alpha_1} \subset A_{\alpha_2}$
- përgjithësisht $\bar{A}_\alpha \neq \overline{A_\alpha}$, përveç rastit kur $\alpha = 0.5$
- A_1 përbën bërthamën e një bashkësie të turbullt A
- A_{0^+} përbën suportin e një bashkësie të turbullt A

2.2 Relacionet e turbullta

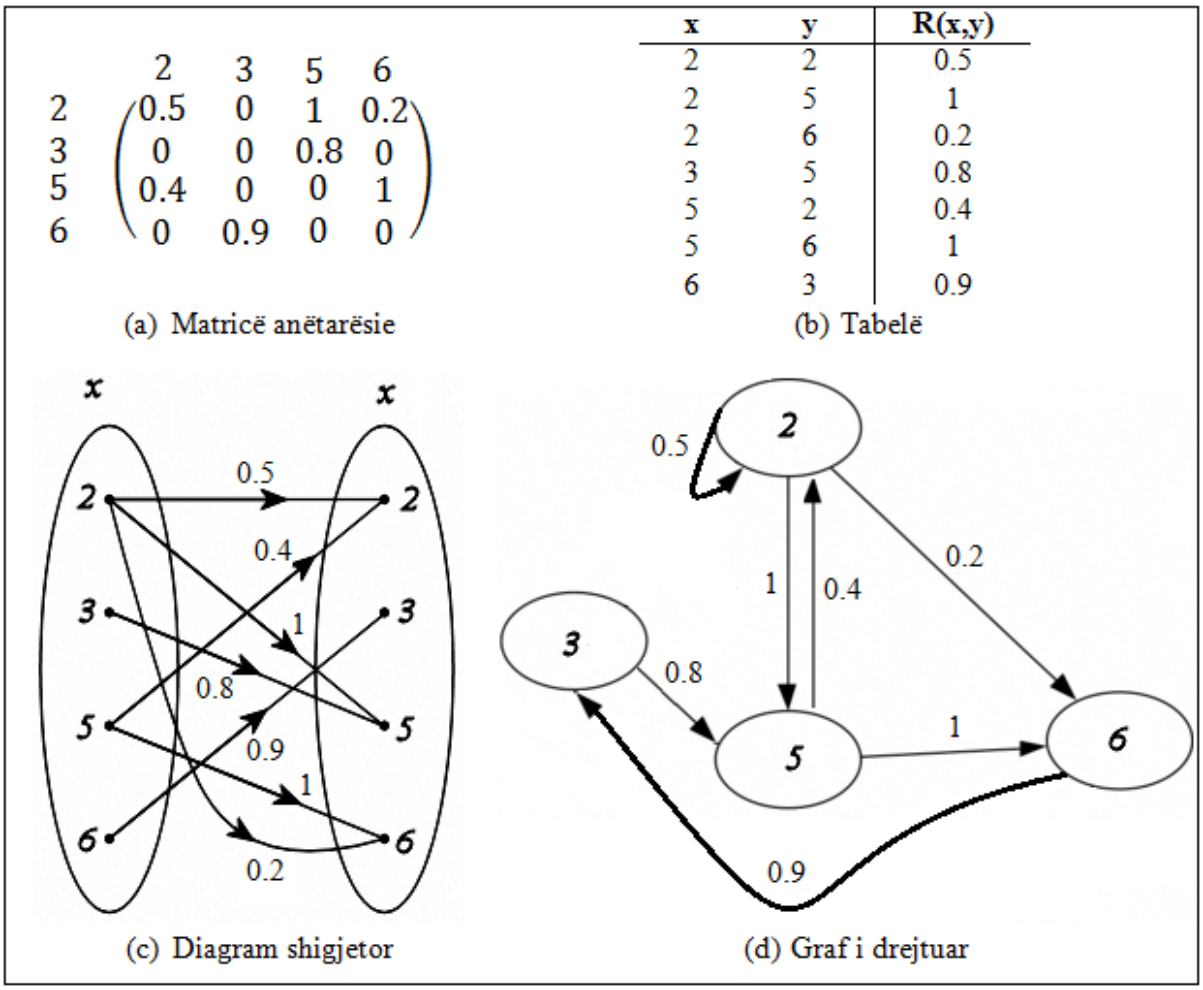
Koncepti i relacionit është thelbësor në trajtimin e një sërë çështjeve të lidhura ngushtë me natyrën aplikative të logjikës së turbullt. Relacionet përfshihen gjerësisht në arsyetimin e përafërt, sistemet e bazuara në rregulla, simulimet jolineare, vlerësimet sintetike, klasifikimet, njohjen e modeleve dhe sistemet e kontrolit. Një relacion klasik paraqet prezencën apo mungesën e shoqërimit, ndërveprimit apo ndërlidhjes ndërmjet dy apo më shumë bashkësive. Ky koncept mund të përgjithësohet duke lejuar shkallë të ndyshme të shoqërimit apo ndërveprimit ndërmjet elementëve, duke formuar kështu relacionet e turbullta. Pra relacionet e turbullta janë një përgjithësim i relacioneve klasike (saktësisht në të njejtën mënyrë siç bashkësitë e turbullta ishin një përgjithësim i bashkësive klasike). Një relacion n-ar midis bashkësive $X_1, X_2, \dots, X_n$ përkufizohet si një nënbashkësi e prodhimit kartezian $X_1 \times X_2 \times \dots \times X_n$ dhe shënohet si $R(X_1, X_2, \dots, X_n)$. Më shumë interes ndër relacionet n-are përbën rasti i relacioneve binare (n=2). Një relacion binar i turbullt pasqyron elementet e një bashkësie, psh A, në elemente të një bashkësie tjetër, psh B, duke e shprehur shkallën e relacionit ndërmjet çifteve të renditura me anë të një funksioni anëtarësie i cili merr vlera në segmentin [0,1] (në një farë mënyre relacionet binare të turbullta krijojnë edhe mundësinë e përgjithësimit të konceptit të funksionit)
Disa nga karakteristikat kryesore të një relacioni të turbullt $R(A, B)$ janë [15]:

- bashkësia e përcaktimit e cila është një bashkësi me elementë nga bashkësia A me funksion anëtarësie $\mu(x) = \max_{y \in B} R(x, y)$
- bashkësia e vlerave e cila është një bashkësi me elementë nga bashkësia B me funksion anëtarësie $\mu(y) = \max_{x \in A} R(x, y)$

- lartësia e cila është një vlerë numerike dhe përkufizohet si $\max_{x\in A} \max_{y\in B} R(x,y)$, pra është vlera më e lartë e anëtarësisë që ka ndonjë çift i relacionit.

Për një relacion binar të turbullt $\mathrm{R(A,B)}$ mund të përkufizojmë relacionin e anasjelltë të tij $\mathrm{R^{-1}(B,A)}$ i cili është një relacion në $B \times A$ i tillë që $\mathrm{R^{-1}(y,x) = R(x,y)}$, $\forall \mathrm{x} \in \mathrm{A}, \forall \mathrm{y} \in \mathrm{B}$. Për dy relacione binare të turbullta që kanë bashkësi të përbashkët si $P(A,B)$ dhe $Q(B,C)$ (pra B është i përbashkët) mund të përkufizojmë relacionin e përbërjes së këtyre relacioneve si $R(A,C) = P(A,B)°Q(B,C)$, ku $R(x,z) = \max_{\mathrm{y\in B}} \min[P(x,y), Q(y,z)]$, $\forall x \in A, \forall z \in C$ [15],[16].

Ndër relacionet e turbullta binare, një fushë të veçantë interesi përbëjnë relacionet mbi një bashkësi të vetme (pra të formës $R(A,A)$). Ekzistojnë disa mënyra për paraqitjen e relacioneve të tilla, si me anë të matricave të anëtarësisë, tabelave, diagrameve shigjetorë dhe grafeve të drejtuar, të cilat i ilustrojmë në figurën 11:



Figurë 11 - Mënyra të paraqitjes së relacioneve të turbullta

Në relacionet e turbullta, ashtu si në bashkësitë e turbullta, mund të kryhen veprimet themelore e mëposhtme [16]:

Veprimi	Shënimi	Funksioni i anëtarësisë
Bashkimi	$R \cup S$	$\mu_{R\cup S}(x,y) = \max(\mu_R(x,y), \mu_S(x,y))$
Prerja	$R \cap S$	$\mu_{R\cap S}(x,y) = \min(\mu_R(x,y), \mu_S(x,y))$
Komplementi	$\bar{R}$	$\mu_{\bar{R}}(x,y) = 1 - \mu_R(x,y)$

Tabelë 2 - Veprimet themelore mbi relacionet e turbullta

Ashtu si për relacionet klasike mbi një bashkësi të vetme, edhe për relacionet e turbullta mbi një bashkësi të vetme përkufizojmë disa veti kryesore (disa prej të cilave në formë më të përgjithësuar) si pasqyrimi, anti-pasqyrimi, simetria, anti-simetria, kalimi si më poshtë [15],[16]:

- Një relacion i turbullt $R(A,A)$ është pasqyrues nëse $R(x,x) = 1, \forall x \in A$
- Një relacion i turbullt $R(A,A)$ është anti-pasqyrues nëse $R(x,x) \neq 1, \forall x \in A$
- Një relacion i turbullt $R(A,A)$ është simetrik nëse $R(x,y) = R(y,x), \forall x \in A, y \in A$
- Një relacion i turbullt $R(A,A)$ është anti-simetrik nëse $R(x,y) = 0$ ose $R(y,x) = 0, \forall x, y \in A, x \neq y$
- Një relacion i turbullt $R(A,A)$ ka vetinë e kalimit nëse $\forall x, y \in A$ plotësohet: $R(x,z) \geq \max_{y\in A} \min(R(x,y), R(y,z))$

Në bazë të vetive të mësipërme për relacionet e turbullta binare mbi një bashkësi dallohen disa lloje kryesore si [17]:

- Ekuivalenca: Një relacion i turbullt R(A,A) është ekuivalencë nëse gëzon vetitë e pasqyrimit, simetrisë dhe kalimit.
- Gati-ekuivalenca: Një relacion i turbullt R(A,A) është gati-ekuivalencë nëse gëzon vetitë e simetrisë dhe kalimit.
- Renditja e pjesshme: Një relacion i turbullt R(A,A) është renditje e pjeshme nëse gëzon vetitë e pasqyrimit, anti-simetrise dhe kalimit.
- Renditja strikte: Një relacion i turbullt R(A,A) është renditje strikte nëse gëzon vetitë e anti-pasqyrimit, anti-simetrisë dhe kalimit.

2.3 Implikimet e turbullta

Implikimet e turbullta janë thelbësore për procesin e arsyetimit të turbullt, ashtu siç implikimet klasike janë thelbësore për procesin e arsyetimit klasik. Në përgjithësi një implikim i turbullt është një funksion $I: [0,1]^2 \to [0,1]$ i cili përcakton vlerën e vërtetësisë $I(x,y)$ të implikimit $p \to q$, ku x dhe y janë përkatësisht vlerat e

vërtetësisë së pohimeve p dhe q. Ekziston një pafundësi mënyrash si mund të shprehet një implikim, më konkretisht konsiderohet si implikim çdo funksion $I: [0,1]^2 \to [0,1]$ që për çdo $x, x_1, x_2, y, y_1, y_2 \in [0,1]$ plotëson kriteret themelore [18],[19]:

(1) Nqs $x_1 \leq x_2$ atëherë $I(x_1, y) \geq I(x_2, y)$ (pra $I(x, y)$ duhet të jetë monoton zbritës në lidhje me variablin x)
(2) Nqs $y_1 \leq y_2$ atëherë $I(x, y_1) \leq I(x, y_2)$ (pra $I(x, y)$ duhet të jetë monoton rritës në lidhje me variablin y)
(3) $I(0, y) = 1$ (dominimi i jo-të vërtetës për variablin x)
(4) $I(x, 1) = 1$ (dominimi i të vërtetës për variablin y)
(5) $I(1,0) = 0$ (vlerat kufitare)

Përveç kritereve themelore të lartpërmenduara, ekzistojnë dhe kritere të tjera plotësuese të cilat janë përdorur me sukses për qëllime optimizimi në skenare të ndryshme aplikative. Disa prej kritereve plotësuese të përdorura më gjerësisht janë:

(6) $I(x, x) = 1$ (identiteti)
(7) $I(1, y) = y$ (neutraliteti i së vërtetës për variablin x)
(8) $I(x, I(y, z)) = I(I(x, y), z)$ (parimi i shkëmbimit)
(9) $I(a, b) = 1$ atëherë dhe vetëm atëherë kur $a \leq b$. (kriteri renditës)

Theksojmë që pavarësisht se ekziston një pafundësi funksionesh që plotësojnë kriteret e dhëna në segmentin e vazhduar [0,1], është vetëm një funksion që i plotëson këto kritere në bashkësinë diskrete {0,1}, i cili është implikimi i logjikës klasike.
Për të gjeneruar implikime të turbullta përdoren tri mënyra kryesore të bazuara në logjikën klasike që janë: implikimet-S, implikimet-R dhe implikimet-QL [20],[21]:

- Një implikim-S gjenerohet nga një t-bashkënormë dhe nga mohimi i turbullt me anë të relacionit: $I(x, y) = S(N(x), y)$
- Një implikim-R gjenerohet me anë të një konjuksioni në [0,1] me anë të relacionit $I(x, y) = \sup\{t \in [0,1] | C(x, t) \leq y\}, \ \forall x, y \in [0,1]$
- Një implikim-QL gjenerohet nga një t-normë, një t-bashkënormë dhe një mohim me anë të relacionit $I(x, y) = S(N(x), T(x, y))$

Përveç këtyre tri mënyrave, gjithashtu ekzistojnë shumë klasa të tjera implikimesh të turbullta të cilat nuk gjenerohen nga operatorët themelorë të logjikës së turbullt si mohimi, konjuksioni (t-normat) apo disjunksioni (t-bashkënormat).
Disa nga implikimet e turbullta më të njohura janë implikimi Reichenbach, implikimi Gödel, implikimi Goguen, implikimi Lukasiewicz, implikimi Kleene-Dienes, implikimi Zadeh etj, të cilët i japim të përmbledhur në tabelën në vijim duke vënë në dukje për secilin prej tyre klasën që i përket, si dhe kriteret plotësuese (përveç kritereve themelore) që përmbushin:

Emri	Forma $I(x,y)$	Klasa	Kritere plotësuese
Reichenbach	$1-x+xy$	S	7, 8
Gödel	$\begin{cases}1 & nqs\ x \le y \\ y & nqs\ x > y\end{cases}$	R	6, 7, 8, 9
Goguen	$\begin{cases}1 & nqs\ x < y \\ y/x & nqs\ x \ge y\end{cases}$	R	6, 7, 8, 9
Lukasiewicz	$\min(1,\ 1-x+y)$	R, S	6, 7, 8, 9
Pseudo-Lukasiewicz	$\min(1, \frac{1-x+(1+\lambda)y}{1+\lambda x})$	R, S	6, 7, 8, 9
Kleene-Dienes	$\max(1-x, y)$	S, QL	7, 8
Zadeh	$\max(1-x, \min(x,y))$	QL	7
Gaines-Rescher	$\begin{cases}1 & nqs\ x \le y \\ 0 & nqs\ x > y\end{cases}$	--	6, 7, 8, 9
Wu	$\begin{cases}1 & nqs\ x \le y \\ \min(1-x,y) & nqs\ x > y\end{cases}$	--	6, 9

Tabelë 3 - Format kryesore të funksioneve implikim të turbullta

2.4 Zgjatime të logjikës së turbullt

Gjatë këtij kapitulli diskutuam rreth logjikës së turbullt si një mjet i rëndësishëm në trajtimin e pasigurisë. Ideja qëndrore ishte në përdorimin e anëtarësiësë së pjesshme (në bashkësi dhe relacione) apo të vërtetësië së pjesshme (në pohimet logjike), pra në vend të vlerave në bashkësinë diskrete {0,1}, të përdoren vlera në segmentin e vazhduar [0,1]. Kjo ide është zgjeruar më tej duke sugjeruar që për të shprehur pasigurinë të mos përdoret një vlerë e vetme në [0,1], por të përdoret një bashkësi vlerash, apo dhe funksione. Në këtë mënyrë janë zhvilluar teori si zgjatime të logjikës së turbullt, ku ndër më të njohurat janë logjika e turbullt intuitive, logjika turbullt e llojit 2, logjika e turbullt bipolare etj. Në këtë seksion do të japim një prezantim të përmbledhur rreth logjikës së turbullt intuitive dhe logjikës së turbullt të llojit 2.

2.4.1 Logjika e turbullt intuitive

Logjika e turbullt intuitive bazohet në teorinë e bashkësive të turbullta intuitive (ashtu si logjika e turbullt bazohet në teorinë e bashkësive të turbullta apo logjika klasike bazohet në teorinë e bashkësive klasike). Koncepti i bashkësisë së turbullt intuitive

është prezantuar nga K. Atanassov si një përgjithësim i konceptit të bashkësisë së turbullt. Elementët e një bashkësie të turbullt intuitive karakterizohen edhe nga vlera anëtarësie, edhe nga vlera jo-anëtarësie. Përdorimi i bashkësive të turbullta intuitive në vend të bashkësive të turbullta prezanton shkallë lirie shtesë (indeksi i jo-anëtarësisë) në përshkrimin e bashkësive [22]. Ky përgjithësim i bashkësive të turbullta na mundëson kapacitete më të larta në trajtimin e pasigurisë, duke bërë që bashkësitë e turbullta intuitive të përdoren me sukses në shumë probleme të botës reale si në përpunimin e imazheve, mjekësinë diagnostikuese, analizën e tregjeve, vendimmarrjen në grup etj.
Një bashkësi e turbullt intuitive përkufizohet si $A = \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$, ku $\mu_A: X \to [0,1]$, $\nu_A: X \to [0,1]$ me kushtin $\mu_A(x) + \nu_A(x) \leq 1, \forall x \in X$. Për çdo element x të bashkësisë universale X, funksioni $\mu_A(x)$ shpreh shkallën e anëtarësisë së tij ndërsa funksioni $\nu_A(x)$ shpreh shkallën e jo-anëtarësisë së tij në bashkësinë e turbullt intuitive A. Gjithashtu për çdo element të një bashkësie të turbullt intuitive përkufizohet indeksi i hezitimit si $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$. Është e qartë se $0 \leq \pi_A(x) \leq 1$ dhe ky indeks shpreh mungesën e informacionit nëse elementi x i përket bashkësisë A apo jo [23].

Edhe për bashkësitë e turbullta intuitive, në bazë të funksionit të anëtarësisë dhe funksionit të jo-anëtarësisë mund të përkufizohen konceptet themelore si barazia, nënbashkësia dhe kardinaliteti [23]:

- Dy bashkësi të turbullta intuitive A dhe B janë të barabarta nëse $\mu_A(x) = \mu_B(x)$ dhe $\nu_A(x) = \nu_B(x),\ \forall x \in X$
- Një bashkësi e turbullt intuitive A është nënbashkësi e një bashkësie të turbullt intuitive B nëse $\mu_A(x) \leq \mu_B(x)$ dhe $\nu_A(x) \geq \nu_B(x),\ \forall x \in X$
- Kardinaliteti (numri i elementëve) i një bashkësie të turbullt intuitive A nuk shprehet si një vlerë e vetme por si një interval $[minCard_A, maxCard_A]$, ku $minCard_A = \sum_{i=1}^{n} \mu_A(x)$ dhe $maxCard_A = \sum_{i=1}^{n} (\mu_A(x) + \pi_A(x))$

Gjithashtu për bashkësitë e turbullta intuitive kanë vend edhe veprimet themelore si prerja, bashkimi dhe komplementi:

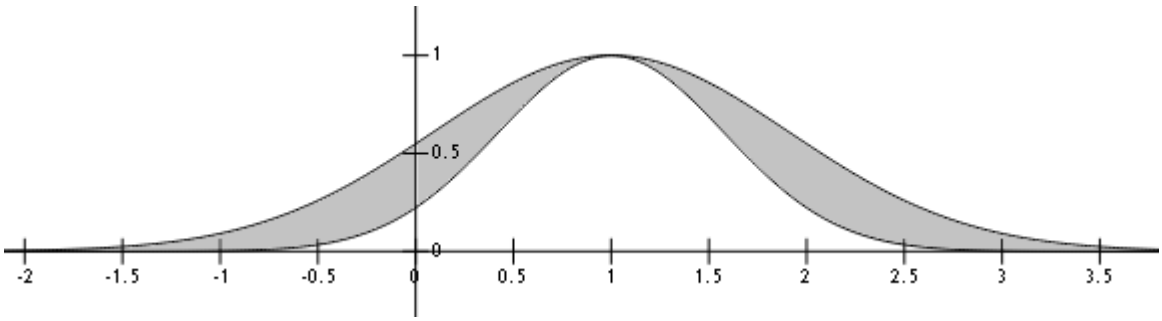
- $A \cap B = \{< x, \min(\mu_A(x), \mu_B(x), \max(\nu_A(x), \nu_B(x)) > | x \in X\}$
- $A \cup B = \{< x, \max(\mu_A(x), \mu_B(x), \min(\nu_A(x), \nu_B(x)) > | x \in X\}$
- $\bar{A} = \{< x, \nu_A(x), \mu_A(x) >\}$

Një nga sfidat e rëndësishme në përdorimin në praktikë të logjikës së turbullt intuitive është ndërtimi i funksioneve të anëtarësisë dhe funksioneve të jo-anëtarësisë. Dy mënyrat kryesore të trajtimit të këtij problemi janë me anë të teknikave të automatizuara si dhe me anë të ekspertizës njerëzore [24]. Modelimi me anë të logjikës së turbullt intuitive është veçanërisht mjaft i përdorshëm në situata ku kemi të

bëjmë me opinione njerëzore apo shprehje idesh në vendimmarrje ku të dhënat janë të formës “Po”, “Jo”, “Nuk e di.”, “Nuk jam i sigurt”, “Nuk kam ndonjë ide.” etj.

2.4.2 Logjika e turbullt e llojit 2

Logjika e turbullt e llojit 2 bazohet në teorinë e bashkësive të turbullta të llojit 2. Një bashkësi e turbullt e llojit 2 është një bashkësi në të cilën ka pasiguri jo vetëm në bashkësinë në tërësi, por edhe në funksion e anëtarësisë të një pjese (ose të gjithë) elementëve të bashkësisë. Pra për një element të bashkësisë vlera e anëtarësisë përgjithësisht nuk është një vlerë konstante, por është një funksion. Pra anëtarësia për një bashkësi të turbullt të llojit 2 nuk paraqitet si një lakore (siç është për bashkësitë e turbullta), por si një zonë (si në figurën 12) [25]. Koncepti i bashkësisë së turbullt të llojit 2 mund të përgjithësohet më tej me bashkësitë e turbullta të llojit 3, të cilat janë bashkësi ku anëtarësia e elementëve paraqitet me anë bashkësive të turbullta të llojit të dytë, apo akoma më gjerë me anë të bashkësive të turbullta të llojit n. Gjithsesi bashkësitë e turbullta të llojeve më të mëdhenj se 2 mbeten kuriozitete matematikore dhe nuk kanë pasur ndonjë zbatueshmëri në praktikë.



Figurë 12 - Shembull i zonës së anëtarësisë për një bashkësi të turbullt të llojit 2

Një bashkësi e turbullt e llojit 2, e shënuar $\check{A}$, karakterizohet nga një funksion anëtarësie i llojit 2: $\mu_{\tilde{A}}(x,u)$ ku $x \in X$ dhe $u \in J_x \subseteq [0,1]$. Më konkretisht:

$$\tilde{A} = \{((x,u), \mu_{\tilde{A}}(x,u)) | \forall x \in X \ \forall u \in J_x \subseteq [0,1]\}$$

ku $0 \leq \mu_{\tilde{A}}(x,u) \leq 1$. Gjithashtu bashkësinë $\tilde{A}$ mund ta shprehim si:

$$\tilde{A} = \sum_{x \in X} \sum_{u \in J_x} \frac{\mu_{\tilde{A}}(x,u)}{(x,u)}, \quad J_x \subseteq [0,1]$$

ku ΣΣ shpreh bashkimin mbi të gjitha x dhe u të pranueshme. Për bashkësitë e vazhdueshme në vend të operatorit ΣΣ, përdoret operatori $\int\int$ [25] .

Ndonëse implementimi i sistemeve bazuar në logjikën e turbullt të llojit 2 paraqet më tepër vështirësi krahasuar më sistemet bazuar në logjikën e turbullt, logjika e turbullt e llojit 2 ka gjetur mjaft zbatime në praktikë në fusha si parashikimi i serive kohore,

klasifikimi i rrymave video të koduara, logjistikë, përpunimi paraprak i imazheve radiografike etj [26]. Në përgjithësi modelimi me anë të logjikës së turbullt të llojit 2 është i përshtatshëm në rrethana ku:

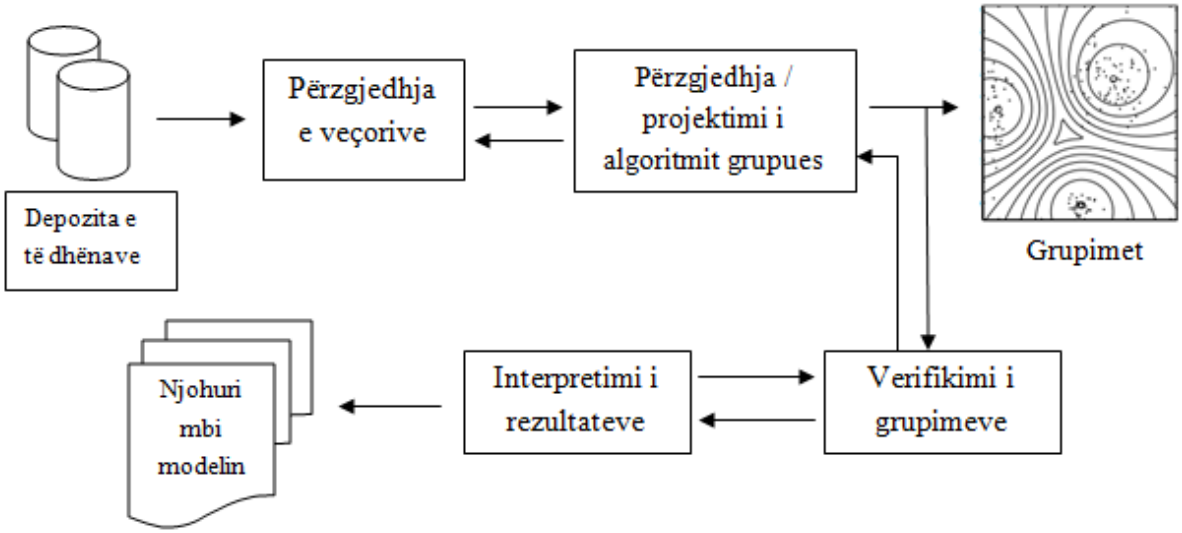
- sistemi gjenerues i të dhënave varion në kohë por përshkrimi matematikor i këtij variacioni nuk dihet
- në proces ka zhurmë dhe përshkrimi matematikor i saj nuk dihet
- veçoritë në një aplikim të analizës së grupimeve kanë atribute statistikore jo-stacionare, përshkrimi matematikor i të cilave nuk dihet
- analizimi i të dhënave nëpërmjet anketimeve ku në pyetjet e anketimit ka fjalë (ose shprehje) që implikojnë mungesë precizioni

3. ANALIZA E GRUPIMEVE

Zhvillimet galopante në teknologjinë e informacionit, veçanërisht në fusha si sistemet e depozitimit të të dhënave, mbledhja e të dhënave nëpërmjet sensorëve automatikë apo nëpërmjet aplikacioneve të panumërta në internet, kanë bërë që gjithnjë e më shumë të jemi të pajisur me volume të jashtëzakonshme të dhënash. Kjo rritje si në volum, ashtu dhe në shumëllojshmërinë e të dhënave ka parashtruar dhe vazhdon të parashtrojë nevojshmëri për metodologji të avancuara të përpunimit, analizimit dhe kuptimit të të dhënave. Metodologjitë e analizimit të të dhënave i klasifikojmë në dy kategori kryesore: në analizën e mbikëqyrur dhe analizën e pambikëqyrur të të dhënave. Teknikat e analizës së mbikëqyrur të të dhënave veprojnë në kushte ku kemi informacion rreth një pjese të konsiderueshme të të dhënave (p.sh. mund të kemi si informacion të dhëna të kategorizuara ose shpërndarje statistikore të të dhënave), të cilin e shfrytëzojmë për të realizuar parashikime (klasifikime/kategorizime) për të dhëna të reja. Nga ana tjetër teknikat e analizës së pambikëqyrur veprojnë në kushte ku nuk ka ndonjë model të specifikuar paraprakisht rreth të dhënave, por synohet të zbulohen struktura të brendshme duke i organizuar të dhënat në grupime bazuar vetëm në ngjashmëritë ndërmjet elementëve individualë. Në lidhje me metodologjitë e lartpërmendura përdoret shpesh termi "mësim" dhe tashmë për to janë standarde emërtimet "mësim i mbikëqyrur" dhe "mësim i pambikëqyrur". Në këtë disertacion fokusi kryesor do të jetë mbi mësimin e pambikëqyrur, por gjithashtu në kapitullin e pestë do të trajtojmë një metodologji hibride si mësimi gjysmë-i mbikëqyrur, që kohët e fundit ka tërhequr masivisht interesin e kërkuesve shkencorë dhe ka gjetur mjaft zbatime të suksesshme.

Analiza e grupimeve është një nga format kryesore të mësimit të pambikëqyrur. Qëllimi thelbësor i analizës së grupimeve është ndarja e të dhënave në grupime (nënbashkësi) në mënyrë të tillë që të dhënat brenda të njëjtit grupim të jenë sa më të ngjashme me njëra-tjetrën dhe sa më të ndryshme nga të dhënat në grupime të tjera. Koncepti i ngashmërisë shprehet në mënyrë sasiore me anë të funksioneve distancë (siç do ta shohim më tej gjatë ketij kapitulli). Për trajtimin e problemit të analizës së grupimeve ekziston një larmi algoritmesh, për të cilët në seksionin vijues do të paraqesim një klasifikim të përgjithshëm. Aftësia e algoritmeve të grupimeve për të vepruar në kushte të mungesës së informacionit i ka bërë ato një mjet mjaft të rëndësishëm në analizën eksploratore të të dhënave dhe kanë gjetur zbatime në një mori disiplinash si biznes, biologji, mjekësi, përpunim imazhesh, psikologji etj [27]. Në aplikimet në biznes analiza e grupimeve ndihmon specalistët e tregjeve dhe grupet e vendim-marrjes në profilizimin e klientëve, në përcaktimin e grupeve të interesit etj.

Në biologji analiza e grupimeve shfrytëzohet në grupimin e gjeneve me karakteristika të ngjashme, në njohjen e modeleve brenda popullatave të gjallesave etj. Në mjekësi me anë të analizës së grupimeve eksplorohen të dhënat e pacientëve të regjistruara nga pajisjet monitoruese, të dhënat e spektometrisë së masës etj.



Figurë 13 – Etapat kryesore të një procesi të analizës së grupimeve

Procesi i analizës së grupimeve kalon nëpër katër etapa kryesore (të cilat janë të ilustruara dhe në figurën e mësipërme) [28]:

- Përzgjedhja e veçorive. Gjatë kësaj etape në morinë e veçorive të të dhënave shumë-përmasore përzgjidhen kandidatët më përfaqësues ose realizohen transformime agreguese të cilat shkrijnë disa veçori në një veçori të vetme përfaqësuese. Një përzgjedhje e mirë ndikon në uljen e volumit përllogaritës duke përmirësuar performancën e të gjithë procedurës së analizës së grupimeve.
- Përzgjedhja/projektimi i algoritmit grupues. Analiza e grupimeve ka gjetur zbatime në një spektër mjaft të gjerë disiplinash dhe për trajtimin e këtij problemi janë zhvilluar një sërë algoritmesh. Ashtu siç do ta demonstrojmë me anë të procedurave eksperimentale gjatë kapitujve të ardhshëm, nuk ekziston një algoritëm që ta trajtojë në mënyrë efiçiente problemin e analizës së grupimeve për të gjitha bashkësitë e të dhënave. Madje për secilin prej algoritmeve ekzistojnë parametra të cilët përcaktohen paraprakisht dhe si rezultat mënyra të ndryshme si mund të operojë i njëjti algoritëm.
- Verifikimi i grupimeve. Për një bashkësi të dhënash një algoritëm grupues gjithnjë gjeneron një rezultat (pavarësisht nëse ky rezultat përfaqëson një

strukturë të brendshme në të dhëna apo jo) dhe përgjithësisht rezultatet e gjeneruar nga algoritme të ndyshme nuk janë të njëjta. Si rrjedhim kriteret vlerësues janë të rëndësishëm për të dhënë informacion mbi cilësinë e grupimeve të gjeneruara.

- Interpretimi i rezultateve. Qëllimi final i analizës së grupimeve është të ofrojë një përshkrim mbi strukturat e brendshme të të dhënave, në mënyrë që të shrytëzohen për zgjidhjen e problemit fillestar (i cili mund të jetë i natyrave nga më të ndryshmet). Interpretimi përfundimtar kryesisht kryhet nga ekspertët e fushave përkatëse, ndonëse shpesh analiza të mëtejshme apo procedura eksperimentale mund të jenë të nevojshme për t'u kryer.

3.1 Klasifikimi i algoritmeve grupues

Problemi i analizës së grupimeve është një problem mjaft i përgjithshëm (për nga shumëllojshmëria e skenareve në të cilat ai përfshihet) dhe për trajtimin këtij problemi janë zhvilluar një mori algorithmesh me përqasje të natyrave të ndryshme. Është i vështirë realizimi i një klasifikimi gjithëpërfshirës për këto algoritme, gjithsesi kategoritë kryesore janë të pranuara gjerësisht. Në këtë seksion do të paraqesim dhe do të vemë në dukje dallimet midis kategorive kryesore si algoritmet hierarkikë përkundrejt algoritmeve particionalë, si dhe algoritmeve të fortë përkundrejt algoritmeve të turbullt (të butë) [29].

Algoritmet hierakikë përkundrejt algoritmeve particionalë: Algoritmet hierakikë të analizës së grupimeve synojnë të ndërtojnë një hieraki grupimesh në një strukturë në formën e një peme (të quajtur "dendroramë"). Për të gjenuar pemën e grupimeve, algoritmet hierarkikë mund të veprojnë në mënyrë ndarëse (nga lart poshtë) ose në mënyrë bashkuese (nga poshtë lart) dhe për të kryer ndarjet/bashkimet bazohen në kritere të cilat shfrytëzojnë funksionet distancë. Nga ana tjetër algoritmet particionalë ofrojnë një ndarje të vetme në grupime. Këto algoritme kanë zbatueshmëri më të lartë veçanërisht kur kemi të bëjmë me bashkësi me numër të madh të dhënash, për të cilat ndërtimi i një dendrograme do të kishte kompleksitet njehsues mjaft të lartë.

Algoritmet klasikë (të fortë) përkundrejt algoritmeve të turbullt (të butë): Algoritmet e fortë të analizës së grupimeve i shpërndajnë të dhënat nëpër grupime ku një e dhënë i përket saktësisht njërit prej grupimeve. Nga ana tjetër algoritmet e turbullt (të butë) i ndajnë të dhënat në grupime ku një e dhënë mund të ketë anëtarësi të pjesshme (një vlerë në segementin [0,1]) në disa grupime njëkohësisht. Ndër algoritmet e turbullt dallojmë më tej në dy nën-kategori të rëndësishme: algoritmet e turbullt probabilistikë dhe algoritmet e turbullt posibilistikë. Algoritmet e turbullt probabilistikë i ndajnë të dhënat në grupime ku një e dhënë ka anëtarësi të pjesshme në disa grupime dhe shuma e këtyre anëtarësive është 1, ndërsa algoritmet e turbullt posibilistikë i ndajnë

të dhënat në grupime ku një e dhënë ka anëtarësi të pjesshme në disa grupime pa pasur ndonjë kufizim. Është e kuptueshme se ndarja probabilistike është një rast i veçantë i ndarjes posibilistike.

3.2 Funksionet e distancës dhe ngjashmërisë

Të gjithë algoritmet e analizës së grupimeve kanë nevojë të shprehin në mënyrë sasiore ngjashmërinë ndërmjet të dhënave (përderisa synimi i tyre ishte ndarja e të dhënave në grupime ku elementët të jenë sa më të ngjashëm me elementët brenda të njëjtit grupim dhe sa më të ndryshëm me elementët e grupimeve të tjera). Për të realizuar shprehjen në mënyrë sasiore të ngjashmërisë shërbejnë funksionet distancë ose funksionet ngjashmëri. Të dy këto lloje funksionesh përdoren për të shprehur ngjashmërinë midis të dhënave, me dallimin se për funksionet distancë sa më e vogël të jetë vlera e funksionit, aq më e madhe është ngjashmëria, ndërsa për funksionet e ngjashmërisë sa më e madhe të jetë vlera e funksionit, aq më e madhe është ngjashmëria. Theksojmë se po diskutojmë për distancë midis elementësh individualë në bashkësinë e të dhënave (ndërsa në seksionin e mëvonshëm rreth algoritmeve hierakike gjithashtu do të zhvillojmë konceptin e distancës midis grupimeve). Gjithashtu ka vend për diskutim edhe për distancë midis një elementi individual dhe një grupimi, por përgjithësisht kjo shprehet si distancë midis elementit në fjalë dhe qendrës së grupimit (pra si distancë midis dy elementësh individualë).

Një funksion distancë është një funksion i formës $d: X \times X \rightarrow R$ i cili gëzon vetitë [30]:

(1) $d(x, y) \geq 0, \forall x, y \in X$
(2) $d(x, y) = d(y, x), \forall x, y \in X$
(3) $d(x, y) = 0$ atëherë dhe vetëm atëherë kur $x = y \; \forall x, y \in X$
(4) $d(x, y) \leq d(x, z) + d(z, y), \; \forall x, y, z \in X$ (mosbarazimi i trekëndëshit)

Nga ana tjetër një funksion ngjashmërie është një funksion s: $X \times X \rightarrow R$ i cili gëzon vetitë [30]:

(1) $0 \leq s(x, y) \leq 1, \forall x, y \in X$
(2) $s(x, y) = s(y, x), \; \forall x, y \in X$
(3) $s(x, y) = 1$ atëherë dhe vetëm atëherë kur $x = y, \; \forall x, y \in X$
(4) $s(x, y) \cdot s(y, z) \leq \big(s(x, y) + s(y, z)\big) s(x, z)$

Gjithashtu ekzistojnë një sërë mënyrash për të përftuar funksion distancë duke u bazuar në funksion ngjashmëri apo anasjelltas. Për të dyja rastet po japim nga një shembull [30]:

- Nëse $s(x,y)$ është një funksion ngjashmëri, atëherë $d(x,y) = 1 - s(x,y)$ është funksion distance.
- Nëse $d(x,y)$ është një funksion distance, atëherë $s(x,y) = e^{-d^2(x,y)}$ është funksion ngjashmëri.

Më poshtë po paraqesim disa nga funksionet distancë ose ngjashmëri për të dhëna numerike, të cilët janë më të përdorur në praktikë. Për secilin rast shprehet distanca ose ngjashmëria midis dy elementëve të të dhënave $U = \{u_1, u_2, \dots, u_n\}$ dhe $V = \{v_1, v_2, \dots, v_n\}$ [31].

- Distanca Euklidiane: është një nga funksionet distancë më të përdorura, përllogaritet me anë të $d(U,V) = \sqrt{\sum_{k=1}^{n}(u_k - v_k)^2}$ Katrori i distancës Euklidiane, pra shprehja $\sum_{k=1}^{n}(u_k - v_k)^2$ nuk është funksion distancë sepse nuk plotëson kriterin e mosbarazimit të trekëndëshit, por përdoret shpesh në algoritme grupues në rastet kur na nevojitet thjesht të krahasojmë vlerat e distancave pa kryer përpunime të mëtejshme.
- Distanca Manhattan (ndryshe quhet dhe distanca e taksisë): shprehet si shuma e vlerave absolute të diferencave përkatëse të veçorive, pra $d(U,V) = \sum_{k=1}^{n} |u_k - v_k|$.
- Distanca maksimale: shprehet si vlera maksimale e diferencës midis dy veçorish respektive në të dhëna, pra $d(U,V) = \max_{k\in\{1,2,\dots,n\}} |u_k - v_k|$.
- Distanca Minkovski: $d(U,V) = \sqrt[m]{\sum_{k=1}^{n}|u_k - v_k|^m}$ është një përgjithësim i distancave të lartpërmendura, konkretisht distanca Euklidiane është rast i veçantë i distancës Minkovski për $m = 2$, distanca Manhattan për $m = 1$, ndërsa distanca maksimale për $m \to \infty$. Këtë distancë do ta kemi subjekt studimi eksperimental sintonizues gjatë kapitullit të katërt.
- Distanca Pearson e korrelacionit: bazohet në koncepte statiskorë dhe përllogaritet si $d(U,V) = \frac{\sum_{k=1}^{n} u_k v_k - n\bar{U}\bar{V}}{(\sum_{k=1}^{n} {u_k}^2 - n\bar{U}^2)(\sum_{k=1}^{n} {v_k}^2 - nV^2)}$, ku $\bar{U} = \frac{\sum_{k=1}^{n} u_k}{n}$ dhe $\bar{V} = \frac{\sum_{k=1}^{n} v_k}{n}$
- Distanca Mahalanobis: përllogaritet si $d(U,V) = (U - V)^T S^{-1}(U - V)$, ku S është matrica e kovariancës së bashkësisë së të dhënave, pra ndryshe nga të gjithë funksionet e kësaj liste, distanca midis dy elementësh nuk varet vetëm prej tyre (por edhe nga bashkësia e të gjithë të dhënave).
- Ngjashmëria e kosinusit: bazohet në formulën e kosinusit të këndit midis dy vektorëve në hapësirën n- përmasore, përllogaritet si:

$$s(U,V) = \frac{\sum_{k=1}^{n} u_k v_k}{(\sum_{k=1}^{n} {u_k}^2)(\sum_{k=1}^{n} {v_k}^2)}$$

- Ngjashmëria e informacionit të përbashkët të normalizuar: bazohet në koncepte të teorisë së informacionit, përllogaritet si $s(U,V) = \frac{I(U,V)}{\sqrt{H(U)H(V)}}$, ku I(U,V) shpreh informacionin e përbashkët, ndërsa $H(U)$ dhe $H(V)$ entropinë e U dhe V.

3.3 Algoritmet hierarkikë

Qëllimi i algoritmeve hierarkikë të grupimit është të gjenerojnë një hierarki grupimesh në një strukturë në formën e një peme (dendrogramë). Struktura e gjeneruar nga këto algoritme është mjaft informuese për bashkësinë e të dhënave, por zbatueshmëria është e kufizuar në bashkësi me numër jo shumë të madh të dhënash (për arsye të kompleksitetit të lartë njehsues që karakterizon këto algoritme). Ekzistojnë dy mënyra kryesore për ndërtimin e strukturës hierarkike në të dhëna nga algoritmet hierarkikë: mënyra bashkuese (nga poshtë lart) dhe mënyra ndarëse (nga lart poshtë) [32]. Në vijim do të diskutojmë në mënyrë të hollësishme mbi mënyrat bashkuese (të cilat janë më të përdorura) dhe më pas do të diskutojmë parimet e përgjithshme të mënyrave ndarëse.

3.3.1 Algoritmet hierarkikë bashkues

Algoritmet hierakikë bashkues në momentin fillestar çdo të dhënë (pra çdo element të bashkësisë së të dhënave) e konsiderojnë si një grupim. Më tej gjatë çdo përsëritjeje (iteracioni) të algoritmit dy nga këto grupime shkrihen me njëri-tjetrin për të formuar një grupim më të madh. Kjo procedurë vazhdon deri sa të gjitha grupimet të shkrihen në një grupim të vetëm të madh (rrënja e pemës) ose në raste të rralla për algoritmin mund të vendosen dhe kritere të tjera përfunduese (në mënyrë të parakohshme, pra pa e ndërtuar hierkinë të plotë). Në çdo përsëritje, përzgjedhja e çiftit të grupimeve të cilët do të shkrihen në një grupim më të madh, realizohet duke shfrytëzuar funksionet distancë midis grupimeve (të cilët do të sqarohen me kujdes në faqen vijuese); më konkretisht në çdo përsëritje të algoritmit përzgjidhet çifti i grupimeve që janë më të afërt (kanë distancën më të vogël) me njëri-tjetrin.

Le të shënojmë bashkësinë e të dhënave $\mathrm{X} = \{\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_\mathrm{n}\}$. Gjithashtu le të jetë përzgjedhur një funksion $d(x,y)$ për të shprehur distancën midis dy elementëve individualë x dhe y, si dhe një funksion $\mathrm{D}(G_i, G_j)$ për të shprehur distancën midis grupimeve G_i dhe G_j të bashkësisë momentale të grupimeve $\Gamma = \{G_1, G_2, \dots, G_k\}$. Atëherë algoritmin e grupimit hierarkik bashkues e përshkruajmë me anë të pseudokodit të dhënë në faqen vijuese [32]:

1. Për çdo të dhënë le të formojmë një grupim (të përbërë nga një element), pra bashkësia fillestare e grupimeve le të jetë $\Gamma = \{G_1, G_2, \dots, G_n\}$, ku $G_i = \{x_i\}$ për çdo $i \in \{1,2, \dots, n\}$.
2. Shënojmë k = n (numri i grupimeve).
3. Llogariten distancat ndërmjet grupimeve, që momentalisht janë $D(G_i, G_j) = d(x_i, x_j)$, meqë çdo grupim përbëhet nga një e dhënë.
4. Në bashkësinë e grupimeve Γ gjejmë çiftin e grupimeve (G_u, G_v) për të cilët distanca midis tyre është më e vogla, pra $D(G_u, G_v) = \min_{P,Q \in \Gamma} D(P, Q)$.
5. I shkrijmë grupimet G_u dhe G_v në një grupim G_w, pra $G_w = G_u \cup G_v$.
6. Heqim G_u dhe G_v nga Γ , si dhe shtojmë G_w në Γ.
7. $k = k - 1$ (numri i grupimeve zvogëlohet me një)
8. Llogaritim distancën midis G_w(grupimit të ri) dhe të gjithë grupimeve të tjera në Γ.
9. Nqs $k > 1$ atëherë vazhdo në hapin 4.
10. FUND

Figurë 14 – Pseusdokodi përshkrues i algoritmeve hierarkikë bashkues

Ekzistojnë disa variante të algoritmit hierakik bashkues, në varësi të mënyrës si shprehet distanca $D(G_i, G_j)$ midis grupimeve. Disa nga variantet kryesore janë algoritmi hierakik i lidhjes njëfishe (minimale), algoritmi hierakik i lidhjes së plotë (maksimale) dhe algoritmi hierarkik i lidhjes mesatare, parimet e punës të së cilëve po i japim në vijim [33]:

- Algoritmi hierakik i lidhjes njëfishe, distancën midis dy grupimesh e shpreh si distanca më e vogël që ekziston midis ndonjë elementi të grupimit të parë dhe ndonjë elementi të grupimit të dytë, pra distanca midis dy grupimesh $G_p$dhe $G_q$shprehet si:
$$D(G_u, G_v) = \min_{x \in G_u, y \in G_v} d(x, y)$$
- Algoritmi hierakik i lidhjes së plotë, distancën midis dy grupimesh e shpreh si distanca më e madhe që ekziston midis ndonjë elementi të grupimit të parë dhe ndonjë elementi të grupimit të dytë, pra distanca midis dy grupimesh $G_p$dhe $G_q$shprehet si:
$$D(G_u, G_v) = \max_{x \in G_u, y \in G_v} d(x, y)$$
- Algoritmi hierakik i lidhjes mesatare, distancën midis dy grupimesh e shpreh si mesataren aritmetike të të gjithë distancave që ekzistojnë midis ndonjë elementi të grupimit të parë dhe ndonjë elementi të grupimit të dytë, pra distanca midis dy grupimesh $G_p$dhe $G_q$shprehet si:

$$D(G_u, G_v) = \frac{1}{|G_u| \cdot |G_v|} \sum_{x \in G_u, y \in G_v} d(x, y)$$

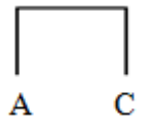
Një aspekt mjaft i rëndësishëm i varianteve të lartpërmendura të algoritmeve hierarkikë bashkues është mundësia për të llogaritur në mënyrë efikase distancën midis një grupimi të ri të sapoformuar nga shkrirja dhe grupimeve të tjera ekzistuese, pra mund të ekzekutojmë në mënyrë efikase hapin e 8-të të algoritmit të dhënë në figurën 14:

- Për algoritmin e lidhjes njëfishe përdorim:
$$D(G_w, P) = \min\{D(G_u, P), D(G_v, P)\}.$$
- Për algoritmin e lidhjes së plotë përdorim:
$$D(G_w, P) = \max\{D(G_u, P), D(G_v, P)\}.$$
- Për algoritmin e lidhjes mesatare përdorim:
$$D(G_w, P) = \frac{|G_u|}{|G_w|} D(G_u, P) + \frac{|G_v|}{|G_w|} D(G_v, P),$$
ku është e qartë se $|G_w| = |G_u| + |G_v|$.

Le të japim një shembull të zbatimit të algoritmit hierarkik të lidhjes njëfishe mbi një bashkësi abstrakte prej pesë elementësh A,B,C,D,E për të cilat distancat midis çdo çifti janë të dhëna paraprakisht në një matricë. Në çdo hap ne përshkruajmë si ndyshon bashkësia e grupimeve (duke përditësuar matricën) dhe ndërtojmë dendrogramën, e cila paraqet vizualisht hierarkinë e grupimeve.

	A	B	C	D	E
A	0	9	3	10	6
B		0	12	4	7
C			0	5	9
D				0	8
E					0

Fillimisht shkrijmë grupimet A dhe C, duke krijuar kështu dendrogramën:

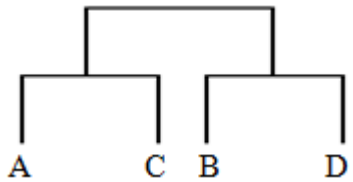


	(A,C)	B	D	E
(A,C)	0	9	5	6
B		0	4	7
D			0	8
E				0

Më pas shkrijmë grupimet B dhe D, dhe dendrograma bëhet:

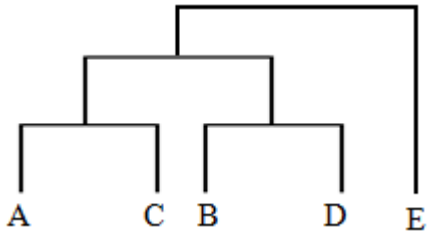
Më pas shkrijmë grupimet (A,C) dhe (B,D) dhe dendrograma bëhet:

	(A,C)	(B,D)	E
(A,C)	0	5	6
(B,D)		0	7
E			0



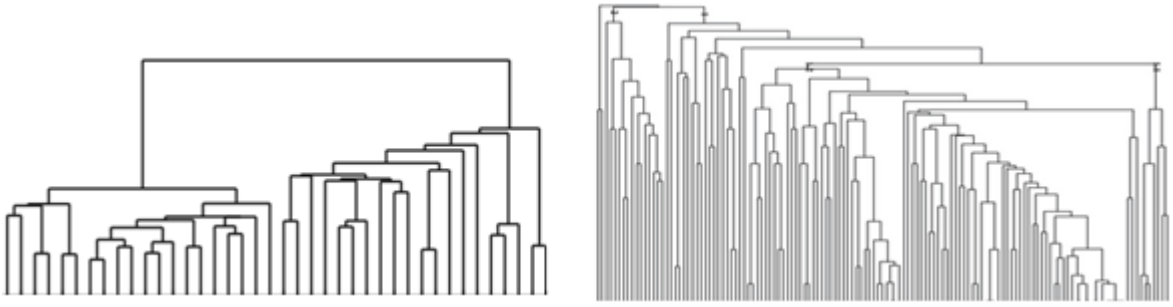
Përfundimisht shkrijmë (A,B,C,D) dhe E, duke përftuar dendrogramën që përfaqëson hierarkinë përfundimtare:

	(A,C,B,D)	E
(A,C,B,D)	0	6
E		0



Figurë 15 - Dendrograma përfundimtare e gjeneruar nga algoritmi hierarkik i lidhjes së njëfishtë

Natyrisht bashkësia e përzgjedhur për ilustrim ishte e vogël. Në figurën e mëposhtme po japim dy shembuj dendrogramesh për bashkësi më të mëdha të dhënash:



Figurë 16 – Shembuj rezultatesh të algoritmeve hierarkikë mbi bashkësi relativisht të mëdha të dhënash të paraqitura vizualisht (me anë të dendrogrameve)

3.3.2 Algoritmet hierarkikë ndarës

Algoritmet hierakikë ndarës (që funksionojnë nga lart poshtë) në momentin fillestar nisin me një grupim të vetëm i cili është e gjithë bashkësia e të dhënave. Ky grupim ndahet në disa nëngrupime bazuar në një nga algoritmet particionalë (të cilët do t'i diskutojmë në seksionin vijues). Më pas kjo procedurë përsëritet në mënyrë rekursive për grupimet e krijuara deri sa të arrijmë në nivelin ku çdo grupim përbëhet nga një element i vetëm. Kjo teknikë është më e ndërlikuar në krahasim me algoritmet hierakikë bashkues, për arsye se për të realizuar ndarjen e grupimeve na nevojitet të vemë në punë ndonjë nga algoritmet particionalë. Nga ana tjetër algoritmet hierakikë ndarës (në krahasim me ato bashkues) kanë si avantazh që nuk janë të detyruar ta gjenerojnë të gjithë hierakinë deri në nivelin më të imët (grupime me një element), por mund të vazhdojnë ekzekutimin deri në një nivel të caktuar paraprakisht [34]. Këto algoritme madje mund të zbatohen edhe mbi bashkësi me numër të madh elemetësh në rastet kur hierarkia nuk do të gjenerohet e plotë, por deri në një nivel të caktuar (me vlerë relativisht të vogël).

3.4 Algoritmet particionalë

Qëllimi i algoritmeve particionalë është të gjenerojnë një ndarje të vetme të dhënave në grupime (dhe jo një hierarki grupimesh). Ndonëse është një problem me kompleksitet njehsues më të vogël në krahasim me grupimet hierakike, këtu ekzistojnë një sërë çështjesh delikate që duhen marrë parasysh si numri i grupimeve, lloji i grupimeve etj. Për sa i përket llojit të grupimeve që mund të gjenerohen, kemi dy lloje kryesore algoritmesh particionalë: klasikë (të fortë) dhe të turbullt ("të butë"). Algoritmet klasikë të analizës së grupimeve i shpërndajnë të dhënat nëpër grupime ku një e dhënë i përket saktësisht njërit prej grupimeve, ndërsa algoritmet e turbullt i ndajnë të dhënat në grupime ku një e dhënë mund të ketë anëtarësi të pjesshme në disa grupime njëkohësisht. Në bashkësinë e algoritmeve të turbullt gjithashtu dallojmë dy nënlloje të rëndësishëm: algoritmet e turbullt probabilistikë, të cilët i ndajnë të dhënat në grupime ku një e dhënë ka anëtarësi të pjesshme në disa grupime me kushtin që shuma e këtyre anëtarësive është 1, si dhe algoritmet e turbullt posibilistikë të cilët i ndajnë të dhënat në grupime ku një e dhënë ka anëtarësi të pjesshme në disa grupime pa pasur ndonjë kufizim. Algoritmet e turbullt (të butë) ofrojnë më tepër mundësi në eksplorimin e të dhënave, duke qenë se për këtë kategori algoritmesh kemi konsiderueshëm më shumë variante të zhvilluar se sa në kategorinë e algoritmeve të fortë (klasikë) [35].

Pavarësisht se algoritmet e grupimit particionalë kanë aplikueshmëri të mirë mbi të dhëna të llojeve të ndryshme, ato dallohen në veçanti për eksplorimin e të dhënave numerike. Gjatë të gjithë nënseksioneve në vijim (përveç rasteve kur është specifikuar ndryshe) bashkësinë e të dhënave e konceptojmë si $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$, ku secili element i saj përmban disa vlera (atribute, përmasa), pra e konceptojmë si një vektor shtyllë me d përmasa ku $\boldsymbol{x_i} = [x_{i1}, x_{i2}, .., x_{id}]^T$, Në vazhdim do të realizojmë një studim sistematik rreth të gjithë nën-kategorive të algoritmeve particionalë që përmendëm.

3.4.1 Algoritmet particionalë klasikë (të fortë)

Algoritmet particionalë klasikë (të fortë) gjenerojnë një ndarje në grupime, ku çdo e dhënë i përket në mënyrë përjashtuese vetëm njërit prej grupimeve. Një ndarje klasike (e fortë) për bashkësinë e të dhënave X është përcaktimi i nënbashkësive jo-boshe $G_1, G_2, \dots, G_k \subset X$ të tilla që të mos ekzistojë ndonjë çift prej tyre që kanë element të përbashkët (pra $G_i \cap G_j = \phi$ për çdo $1 \leq i < j \leq k$) si dhe bashkimi i të gjithë këtyre nënbashkësive të japë bashkësinë X (pra $\bigcup_{i=1}^{k} G_i = X$).

Duke shfrytëzuar konceptin e funksioneve të anëtarësië, një ndarje për bashkësinë X mund të shprehet si një matricë particion $U = [\mu_{ij}]_{k \times n}$ (pra me k rreshta dhe n

shtylla), ku μ_{ij} shpreh anëtarësinë e elementit x_j në grupimin G_i. Atëherë është e qartë se elementët e matricës U duhet të plotësojnë kushtet:

- $\mu_{ij} \in \{0,1\}$ për çdo $i \in \{1,2,\dots,k\}$, $j \in \{1,2,\dots,n\}$ (vlerat e anëtarësisë janë 0 ose 1)
- $\sum_{i=1}^{k} \mu_{ij} = 1$ për çdo $j \in \{1,2,\dots,n\}$ (çdo element i përket një dhe vetëm një grupimi)
- $\sum_{j=1}^{n} \mu_{ij} > 0$ për çdo $i \in \{1,2,\dots,k\}$ (çdo grupim duhet të përmbajë të paktën një element)

3.4.1.1 Algoritmi klasik i mesoreve

Algoritmi klasik i mesoreve është një nga algoritmet më të përdorur. Ai e parashtron problemin e analizës së grupimeve si një problem optimizimi, ku synohet të minimizohet shuma e katrorëve të distancave të secilit element të bashkësisë së të dhënave nga qendra e grupimit ku ai vendoset. Formalisht këtë e shprehim me anë të një funksioni objektiv, të cilin synojmë ta minimizojmë [36]:

$$J(X,U,C) = \sum_{i=1}^{k}\sum_{j=1}^{n} \mu_{ij} d^2(x_j, c_i) = \sum_{i=1}^{k}\sum_{x_j \in G_i} d^2(x_j, c_i)$$

ku X është bashkësia e të dhënave, U është matrica particion (me vlera binare: 0 ose 1) dhe $C = [c_1, c_2, \dots, c_k]$ është një vektor që shpreh qendrat e grupimeve $G_1, G_2, \dots, G_k$. Problemi i optimizimit që është parashtruar, është mjaft kompleks dhe optimizimi global i tij ka një shkallë kompleksiteti tepër të lartë. Përqasja që algoritmi i fortë i mesoreve ofron është një skemë iterative optimizimi e cila konvergjon në optimume lokalë, rrjedhimisht dhe rezultati i ofruar është i mirë, por nuk është e thënë të jetë më i miri i mundshëm. Për përmirësimin e cilësisë së grupimeve mund të përdoren teknika më të avacuara, ku paraprakisht mund të përmendim zbatimin e algoritmit disa herë dhe më pas përzgjedhjen e ndarjes më të mirë me anë të zëvendësimit të vlerave në funksionin objektiv.

Ekzistojnë tre parametra që është e nevojshme të përcaktohen paraprakisht për algoritmin:

- Numri i grupimeve
- Funksioni distancë (ku variantet kryesore janë diskutuar në seksionin 3.2)
- Vlerat fillestare të qendrave të grupimeve (pra vlerat fillestare të vektorit C)

Përveç tre parametrave themelore të lartpërmendura, shpeshherë përcaktohet dhe një parametër shtesë si numri maksimal i përsëritjeve të algoritmit (MAX_IT), i cili mundëson ndërprerjen e algoritmit nëse rezutati i kërkuar nuk gjendet brenda këtij kufiri për numrin e përsëritjeve. Hapat e algoritmit përshkruhen nga pseudokodi:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. Llogaritet distanca ndërmjet secilit element dhe secilës qendër të grupimeve sipas funksionit distancë.
3. Caktojmë çdo element në grupimin me të cilin ka distancën më të vogël nga qendra e tij, pra $\mu_{ij} = \begin{cases} 1, & nqs\ d(x_j, c_i) = \min_{t \in \{1,2,\dots,k\}} d(x_j, c_t) \\ 0, & per\ rastet\ e\ tjera \end{cases}$
4. Llogariten qendrat e reja të grupimeve: $c_i = \frac{\sum_{j=1}^{n} \mu_{ij} x_j}{\sum_{j=1}^{n} \mu_{ij}}$ për çdo $i \in \{1,2,\dots,k\}$
5. $nr = nr + 1$ (rritet me një numri i iteracioneve)
6. Nqs pati ndryshime në matricën particion (gjatë hapit 3) dhe $nr < MAX_IT$ atëherë vazhdo në hapin 2.
7. FUND

Figurë 17 - Pseudokodi përshkrues i algoritmit të fortë të mesoreve

3.4.2 Algoritmet particionalë të turbullt (të butë)

Algoritmet particionalë të turbullt (të butë) gjenerojnë gjithashtu një ndarje të vetme në grupime, por me dallimin se elementët e bashkësisë së të dhënave kanë anëtarësi të pjesshme (një vlerë në segmentin [0,1]) në disa grupime njëkohësisht.Siç përmendëm dhe më sipër, një ndarje e turbullt mund të jetë probabilistike (pra me kufizimin se shuma e anëtarësive të çdo elementi në të gjitha grupimet duhet të jetë 1) ose posibilistike (ku nuk ka ndonjë kufizim). Para se të diskutojmë konkretisht variante algoritmesh të turbullt particionalë, le të përkufizojmë formalisht konceptet e ndarjes probabilistike dhe ndarjes probabilistike.

Një ndarje e turbullt probabilistike mund përshkruhet me anë të matricës particion $U = [\mu_{ij}]_{k \times n}$ (pra me k rreshta dhe n shtylla), ku μ_{ij} shpreh anëtarësinë e elementit x_j në grupimin G_i.Në këtë rast elementët e matricës U duhet të plotësojnë kushtet:

- $\mu_{ij} \in [0,1]$ për çdo $i \in \{1,2,\dots,k\}$, $j \in \{1,2,\dots,n\}$ (vlerat e anëtarësisë janë numra realë midis 0 dhe 1)
- $\sum_{i=1}^{k} \mu_{ij} = 1$ për çdo $j \in \{1,2,\dots,n\}$ (për çdo element shuma e anëtarësisë në grupime duhet të jetë 1)
- $\sum_{j=1}^{n} \mu_{ij} > 0$ për çdo $i \in \{1,2,\dots,k\}$ (çdo grupim duhet të përmbajë të paktën një element)

Një ndarje e turbullt posibilistike mund përshkruhet me anë të matricës particion $U = [\mu_{ij}]_{k\times n}$ (pra me k rreshta dhe n shtylla), ku μ_{ij} shpreh anëtarësinë e elementit x_j në grupimin G_i.Në këtë rast elementët e matricës U duhet të plotësojnë kushtet:

- $\mu_{ij} \in [0,1]$ për çdo $i \in \{1,2,\dots,k\}$, $j \in \{1,2,\dots,n\}$ (vlerat e anëtarësisë janë numra realë midis 0 dhe 1)
- $\sum_{i=1}^{k} \mu_{ij} > 0$ për çdo $j \in \{1,2,\dots,n\}$ (një element nuk mund të ketë anëtarësi 0 në të gjitha grupimet)
- $\sum_{j=1}^{n} \mu_{ij} > 0$ për çdo $i \in \{1,2,\dots,k\}$ (grupimet nuk duhet të jenë boshe)

3.4.2.1 Algoritmi i turbullt i mesoreve

Algoritmi i turbullt i mesoreve, i projektuar nga Dunn dhe i përmirësuar nga Bezdek, është algoritmi i turbullt më të rëndësishëm dhe një nga algoritmet më të spikatur ndër të gjithë algoritmet e mësimit të pambikëqyrur [30]. Edhe ky algoritëm e parashtron problemin analizës së grupimeve si një problem optimizimi, ku synojmë të minimizojmë një funksion objektiv, i cili shpreh shumën e katrorëve të elementëve nga qendrat e grupimeve të shumëzuar me një koeficient i cili është një fuqi e anëtarësisë së elementit në grupimin përkatës. Formalisht funksionin objektiv e shprehim si [37]:

$$J(X,U,C) = \sum_{i=1}^{k}\sum_{j=1}^{n} {\mu_{ij}}^{\varphi} d^2(x_j, c_i)$$

ku X është bashkësia e të dhënave, U është matrica particion (me vlera reale në segmentin $[0,1]$), $C = [c_1, c_2, \dots, c_k]$ është një vektor që shpreh qendrat e grupimeve $G_1, G_2, \dots, G_k$, si dhe φ është koeficienti i turbullsisë (një vlerë reale në $[1,\infty[$).
Përqasja që algoritmi i turbullt i mesoreve ofron është një skemë iterative optimizimi e cila konvergjon në optimume lokale, pra rezultati i ofruar përgjithësisht është i mirë, por nuk është e thënë të jetë më i miri i mundshëm. Ekzistojnë katër parametra që është e nevojshme të përcaktohen paraprakisht për algoritmin:

- Numri i grupimeve
- Funksioni distancë (ku variantet kryesore janë diskutuar në seksionin 3.2)
- Vlerat fillestare të qendrave të grupimeve (pra vlerat fillestare të vektorit C)
- Vlera e koeficientit të turbullsisë φ (një vlerë reale në $[1,\infty[$).

Përveç katër parametrave themelorë të lartpërmendur, është e nevojshme të të përcaktohet dhe një parametër i cili do të kontrollojë mbarimin e algoritmit. Ky parametër zakonisht është shkalla e tolerancës ε, në raste më të rralla përdoret dhe numri maksimal i përsëritjeve të algoritmit (MAX_IT) ose të dy njëkohësisht (edhe ε, edhe MAX_IT). Hapat e algoritmit përshkruhen nga pseudokodi:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogaritet distanca ndërmjet secilit element dhe secilës qendër të grupimeve sipas funksionit distancë.
4. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas $\mu_{ij} = \frac{\left(d(x_j, c_i)\right)^{-\frac{2}{\varphi-1}}}{\sum_{t=1}^{k}\left(d(x_j, c_t)\right)^{-\frac{2}{\varphi-1}}}$
5. Llogariten qendrat e reja të grupimeve: $c_i = \frac{\sum_{j=1}^{n} \mu_{ij}{}^{\varphi} x_j}{\sum_{j=1}^{n} \mu_{ij}{}^{\varphi}}$ për çdo $i \in \{1,2, \dots, k\}$
6. $nr = nr + 1$ (rritet me një numri i iteracioneve)
7. Nqs $\left||U_{nr} - U_{nr-1}|\right| > \varepsilon$ atëherë vazhdo në hapin 2.
8. FUND

Figurë 18 - Pseudokodi përshkrues i algoritmit të turbullt të mesoreve

3.4.2.2 Algoritmi Gustafson-Kessel

Algoritmi i turbullt i mesoreve funksionon në mënyrë mjaft efiçiente nëse forma e grupimeve në hapësirën shumë-dimensionale është hipersferike, kjo për shkak të metrikave të distancës që përdoren në këtë algoritëm. Algoritmi Gustafson-Kessel vjen si një zgjatim i algoritmit të grupimit të turbullt të mesoreve duke ndryshuar metrikën e distancës së përdorur [38]. Në këtë rast shfrytëzohet një metrikë distance adaptive për të mundësuar dallimin e grupimeve në forma të ndryshme gjeometrike pa i ndryshuar përmasat e grupimeve për një bashkësi të dhënash. Funksioni objektiv dhe parametrat për algoritmin Gustafson-Kessel janë njësoj si në algoritmin e turbullt të mesoreve, pra:

$$J(X, U, C) = \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}{}^{\varphi} d^2(x_j, c_i)$$

ku X është bashkësia e të dhënave, U është matrica particion (me vlera reale në segmentin $[0, 1]$), $C = [c_1, c_2, \dots, c_k]$ është një vektor që shpreh qendrat e grupimeve $G_1, G_2, \dots, G_k$, si dhe φ është koeficienti i turbullsisë (një vlerë reale në $[1, \infty[$).
Ndryshimi thelbësor qëndron në mënyrën si shprehet distanca, e cila përllogaritet si:

$$d^2(x_j, c_i) = (x_j - c_i)^T V_i (x_j - c_i)$$

ku $V_i = |\rho_i F_i|^{1/p} F_i^{-1}$, dhe $F_i = \frac{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{\varphi}(x_j-c_i)^T(x_j-c_i)}{\sum_{i=1}^{c}\sum_{j=1}^{n}\mu_{ij}^{\varphi}}$

F_i është matrica kovariancë e grupimit të i-të, pra çdo grupim ka matricën e tij që përdoret në përllogaritjen e distancave duke bërë që çdo grupim ta përshtasë funksionin distancë me strukturën topologjike lokale të të dhënave [39]. Edhe për këtë algoritëm janë të nevojshëm disa parametra paraprakë si numri i grupimeve, vlera e koeficientit të turbullsisë, shkalla e tolerancës dhe matrica fillestare e distancës. Hapat e algoritmit përshkruhen nga pseudokodi:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogaritet matrica e kovariancës: $F_i = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}(x_i-c_j)(x_i-c_j)^T}{\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}}$
4. Llogaritet distanca ndërmjet secilës pikë dhe secilës qendër:
$$d^2(x_j, c_i) = (x_j - c_i)^T V_i (x_j - c_i)\text{ , ku } V_i = |F_i|^{1/p} F_i^{-1}$$
5. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas: $\mu_{ij} = \frac{\left(d(x_j,c_i)\right)^{-\frac{2}{\varphi-1}}}{\sum_{t=1}^{k}\left(d(x_j,c_t)\right)^{-\frac{2}{\varphi-1}}}$
6. Rillogariten qendrat e reja $c_i = \frac{\sum_{j=1}^{n}\mu_{ij}^{\varphi}X_j}{\sum_{j=1}^{n}\mu_{ij}^{\varphi}}$ për çdo $i \in \{1,2,\dots,k\}$
7. $nr = nr + 1$ (rritet me një numri i iteracioneve)
8. Nqs $||U_{nr} - U_{nr-1}|| \geq \varepsilon$ atëherë vazhdo në hapin 2.
9. FUND

Figurë 19 - Pseudokodi përshkrues i algoritmit Gustafson-Kessel

3.4.2.3 Algoritmi Gath-Geva

Algoritmi Gath-Geva është një zgjatim i algoritmit Gustafson-Kessel duke marrë në konsideratë përmasën dhe dendësinë e grupimeve, duke bërë që të ketë efiçiencë më të mirë në rastet kur kemi të bëjmë me veçori të çrregullta [40]. Distanca në këtë rast shprehet si:

$$d^2(x_j, c_i) = (1/p_i)(2\pi)^{n/2}\sqrt{|F_i|}e^{(x_j-c_i)^T V_i (x_j-c_i)}$$

Në relacionin e mësipërm $p_i = \frac{1}{n}\sum_{j=i}^{n} \mu_{ij}$. Nën këndvështrimin e teorisë së probabilitetit, distanca $d^2(x_i, c_j)$ është në përpjesëtim të zhdrejtë me probabilitetin që elementi x_i të ndodhet në grupimin c_j, ku të dhënat supozohen se i përkasin një shpërndarjeje normale. Në krahasim me distancën e algoritmit Gustafson-Kessel, vemë në dukje praninë e një termi eksponencial i cili bën që distanca e algoritmit Gath-Geva të ndryshojë më shpejt [37].

Avantazhi i këtij algoritmi qëndron në faktin se është në gjendje të dallojë grupime të formave, përmasave dhe dendësive të ndryshme. Nga ana tjetër algoritmi ka si disavantazh që i nevojiten vlera fillestare të përafruara mirë, duke qenë së norma eksponenciale e distancës bën që algoritmi të konvergjojë shpejt në një optimum lokal. Zakonisht për të shmangur këtë disavantazh kritik të algoritmit Gath-Geva, vihet në punë një algoritëm tjetër grupimi posaçërisht për të përdorur qendrat e grupimeve që ai do të gjenerojë, si vlera fillestare të grupimeve për algoritmin Gath-Geva. Edhe për këtë algoritëm janë të nevojshëm disa parametra paraprakë të cilat janë të njejtë si në algoritmin Gustafson-Kessel. Hapat e algoritmit përshkruhen nga pseudokodi:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve).
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogaritet matrica e kovariancës: $\sum_i = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}(x_i-c_j)(x_i-c_j)^T}{\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}}$
4. Llogaritet distanca ndërmjet secilës pikë dhe secilës qendër:

   $d^2(x_i, c_j) = (1/p_i)(2\pi)^{n/2}\sqrt{|F_i|}e^{(x_i-c_j)^T V_i (x_i-c_j)}$ ku $V_i = |F_i|^{1/p}F_i^{-1}$ dhe $p_i = \frac{1}{n}\sum_{j=i}^{n}\mu_{ij}$.
5. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas: $\mu_{ij} = \frac{\left(d(x_j,c_i)\right)^{-\frac{2}{\varphi-1}}}{\sum_{t=1}^{k}\left(d(x_j,c_t)\right)^{-\frac{2}{\varphi-1}}}$
6. Rillogariten qendrat e reja $c_i = \frac{\sum_{j=1}^{n}\mu_{ij}^{\varphi}X_j}{\sum_{j=1}^{n}\mu_{ij}^{\varphi}}$ për çdo $i \in \{1,2,\dots,k\}$
7. $nr = nr + 1$ (rritet me një numri i iteracioneve)
8. Nqs $||U_{nr} - U_{nr-1}|| \geq \varepsilon$ atëherë vazhdo në hapin 2.
9. FUND

Figurë 20 - Pseudokodi përshkrues i algoritmit Gath-Geva

3.4.2.4 Algoritmi i grupimit të turbullt bazuar në bërthama

Algoritmi i grupimit të turbullt bazuar në bërthama është një zgjatim i algoritmit të turbullt të mesoreve. Ideja thelbësore e algoritmit është përdorimi i një pasqyrimi jo-linear (i quajtur ndryshe funksioni bërthamë) nga hapësira e veçorive te një hapësirë bërthamë me më shumë dimensione. Siç e përmendëm më parë, algoritmi i grupimit të turbullt të mesore funksionon me efiçiencë në dallimin e grupimeve hipersferike, por ka vështirësi të theksuara në dallimin e grupive me forma të tjera si dhe në dallimin e grupimeve kur ka sasi të konsiderueshme mbivendosjeje midis tyre. Pasqyrimi jo-linear nga hapësira e veçorive te një hapësirë bërthamë me më shumë dimensione bën që strukturat komplese, linearisht të pandashme, të transformohen te hapësira bërthamë në struktura më të thjeshta dhe linearisht të ndashme. Një avantazh tjetër i algoritmit të grupimit të turbullt bazuar në bërthamë është që në rast se numri i grupimeve nuk është parapërcaktuar, bazuar në kritere të ndryshme algortimi mund ta përshtasë vetë këtë numër të grupimeve. Disa nga variantet më të përdorura të funksionit bërthamë janë [41]:

1. Bërthama Gausiane: $K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}}$, ku $\sigma \in R$
2. Bërthama me bazë radiale: $K(x_i, x_j) = e^{-\frac{\sum |x_i^a - x_j^a|^b}{2\sigma^2}}$, ku $a, b > 0$
3. Bërthama hipertangjente: $K(x_i, x_j) = 1 - \tanh \frac{\|x_i - x_j\|^2}{\sigma^2}$

Pasqyrimin jo-linear e shënojmë si: $\Phi: x \to \Phi(x)$. Funksioni objektiv që synojmë të minimizojmë është:

$$J(X, U, C) = \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}^{\varphi} \left\|\Phi(x_j) - \Phi(c_i)\right\|^2$$

dhe katrorin e distancës e përllogarisim si $\left\|\Phi(x_j) - \Phi(c_i)\right\|^2 = K(x_j, x_j) + K(c_i, c_i) - 2K(x_j, c_i)$ ku $K(x, y)$ është funksioni bëthamë. Duke qenë se funksionet kryesore bërthamë të lartpërmendura plotësojnë vetinë $K(x, x) = 1$, atëherë katrorin e distancës e shkruajmë $\left\|\Phi(x_i) - \Phi(c_j)\right\|^2 = 2\left(1 - K(x_i, c_j)\right)$, si dhe funksionin objektiv e rishkruajmë si [42]:

$$J(X, U, C) = 2 \sum_{i=1}^{k} \sum_{j=1}^{n} \mu_{ij}^{\varphi} \left(1 - K(x_i, c_j)\right)$$

Edhe për këtë algoritëm janë të nevojshëm disa parametra paraprakë si numri i grupimeve, koeficienti i turbullsisë, funksioni bërhtamë, vlerat fillestare të qendrave të grupimeve, shkalla e tolerancës, etj. Hapat e algoritmit përshkruhen nga pseudokodi:

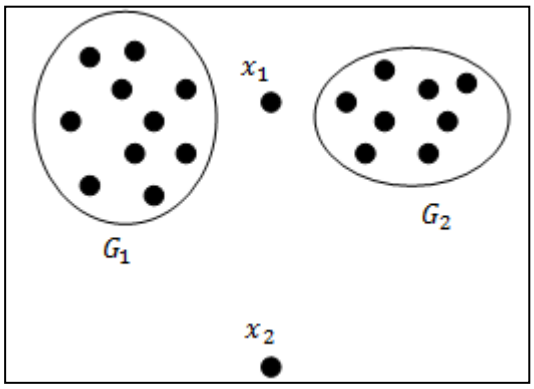
1. Shënojmë $nr = 1$ (numëruesi i iteracioneve).
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogariten qendrat e reja të grupimeve $c_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{\varphi} K(x_j, c_i) x_j}{\sum_{j=1}^{n} \mu_{ij}^{\varphi} K(x_j, c_i)}$

   Përdistësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas $\mu_{ij} = \frac{(1/(1-K(x_j,c_i)))^{1/(\varphi-1)}}{\sum_{t=1}^{k}(1/(1-K(x_j,c_t)))^{1/(\varphi-1)}}$
4. $nr = nr + 1$ (rritet me një numri i iteracioneve)
5. Nqs $\left||U_{nr} - U_{nr-1}|\right| \geq \varepsilon$ atëherë vazhdo në hapin 2.
6. FUND

Figurë 21- Pseudokodi përshkrues i algoritmit të grupimit të turbullt të bazuar në bërthama

3.4.2.5 Algoritmi posibilistik i mesoreve

Të gjitha algoritmet e turbullta të grupimit të përmendura më sipër janë probabilistikë, pra për çdo element të bashkësisë së të dhënave shuma e anëtarësive në grupime është e barabartë me një. Në disa raste, për disa të dhëna kjo shpërndarje nuk është informuese. Bëhet fjalë për të dhëna që janë vlera të largëta (vlera që i përkasin bashkësisë së të dhënave por janë të largëta nga vlerat e tjera) ose për vlerat me zhurmë (vlera të cilat nuk i përkasin bashkësisë së të dhënave, por ndodhen gabimisht, psh matje e gabuar gjatë një eksperimenti etj). Për këto të dhëna shpërndarja probabilistike nuk do ishte e natyrshme. P.sh. marrim në konsideratë bashkësinë e të dhënave në figurën vijuese :



Figurë 22 - Rrethana ku shpërndarja probabilistike nuk do ishte e përshtatshme për elementin x_2

Siç shihet në bashkësinë e të dhënave të figurës 22 janë përcaktuar dy grupime G_1 dhe G_2ku elementët x_1 dhe x_2 nuk kanë anëtarësi të plotë në ndonjërin prej tyre. Këto elementë janë pothuajse të barazlarguar nga qendrat e grupimeve, si rrjedhim për të dy këto elementë anëtarësia në G_1 dhe G_2do të kishte vlerën 0.5 Ndërkohë që ky interpretim konsiderohet i përshtatshëm për elementin x_1, për elementin x_2 nuk konsiderohet i tillë. Ky element vërtet është i barazlarguar nga qendrat e grupimeve, por është larg prej tyre. Në një ndarje posibilistike vlerat e anëtarësisë për elementin x_2 në grupimet G_1 dhe G_2 mund të ishin 0.3 (pra hiqet kufizimi se shuma e anëtarësisë se çdo elementi në grupime duhet të jetë një). Pra në ndarjen posibilistike vlerat e anëtarësisë për vlerat e largëta dhe vlerat me zhurmë janë më përfaqësuese sesa në ndarjen probabilistike.

Algoritmi posibilistik i mesoreve së pari modifikon funksionin objektiv që synon të minimizohet. Natyrisht që funksioni objektiv nuk mund të jetë si te algoritmi i turbullt i mesoreve, pasi në këtë rast ai do të minimizohej duke i bërë të gjitha anëtarësitë me vlera shumë afër 0 (nuk mund t'i bëjmë saktësisht 0 sepse do të thyenim kriterin e ndarjes posibilistike që një element nuk mund të ketë anëtarësi 0 në të gjitha grupimet). Modifikimi i funksionit objektiv realizohet duke shtuar një term i cili pengon vlerat e anëtarësisë të konvergjojnë të gjitha drejt vlerave zero. Pra funksioni objektiv shprehet si shumë e dy termave ku termi i parë tenton që t'i afrojë vlerat e anëtarësisë drejt vlerës zero, ndërsa termi i dytë synon t'i afrojë vlerat e anëtarëisë drejt vlerës një. Më konkretisht funksioni objektiv paraqitet si [43]:

$$J(X,U,C) = \sum_{i=1}^{k}\sum_{j=1}^{n} {\mu_{ij}}^{\varphi} d^2(x_j, c_i) + \sum_{i=1}^{k} \eta_i \sum_{j=1}^{n} \left(1-\mu_{ij}\right)^{\varphi}$$

ku koeficienti $\eta_i > 0$ është një parametër i cili karakterizon grupimin G_1për çdo $i \in \{0,1,\dots,k\}$. Këtë rol mund ta kryejë çdo numër pozitiv, gjithsesi përgjithësisht rekomandohet të përllogaritet si: $\eta_i = \frac{\sum_{j=1}^{n} {\mu_{ij}}^{\varphi} d(x_j,c_i)}{\sum_{j=1}^{n} {\mu_{ij}}^{\varphi}}$.

Në këtë rast vlera e parametrit η_i shpreh një madhësi të rëndësishme përshkruese për grupimin : distanca për të cilën një element ka anëtarësi 0.5 në grupim. Në varësi të formave të grupimeve mund të japim interpretime të ndryshme gjeometrike për parametrin η_i, për shembull në rastin e grupimeve hipersferike shpreh katrorin e diametrit mesatar të grupimit.

Edhe për këtë algoritëm janë të nevojshëm disa parametra paraprakë si numri i grupimeve, koeficienti i turbullsisë, funksioni distancë, vlerat fillestare të qendrave të grupimeve, shkalla e tolerancës etj. Hapat e algoritmit përshkruhen nga pseudokodi:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogaritet distanca ndërmjet secilit element dhe secilës qendër të grupimeve sipas funksionit distancë.
4. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas $\mu_{ij} = \frac{1}{1+\left(\frac{d^2(x_j, c_i)}{\eta_i}\right)^{\frac{1}{\varphi-1}}}$
5. Rillogariten qendrat e reja të grupimeve sipas: $c_i = \frac{\sum_{j=1}^{n} \mu_{ij}{}^{\varphi} x_j}{\sum_{j=1}^{n} \mu_{ij}{}^{\varphi}}$
6. $nr = nr + 1$ (rritet me një numri i iteracioneve)
7. Nqs $\left||U_{nr} - U_{nr-1}|\right| \geq \varepsilon$ atëherë vazhdo në hapin 3.
8. FUND

Figurë 23 - Pseudokodi përshkrues i algoritmit posibilistik të mesoreve

4. TEKNIKA SINTONIZUESE DHE VLERËSUESE, KRAHASIME EKSPERIMENTALE PËR ALGORITMET E GRUPIMIT.

Gjatë kapitullit të kaluar pamë një sërë algoritmesh të zhvilluar për të trajtuar problemin e analizës së grupimeve, madje për secilin prej algoritmeve vumë në dukje se ekzistojnë disa parametra si funksioni distancë, numri i grupimeve, koeficienti i turbullsisë etj, ku për secilin parametër ekzistojnë disa variante se si mund të përzgjidhet. Duke qenë se analiza e grupimeve vepron në rrethana ku nuk kemi informacion rreth strukturës së brendshme të bashkësisë së të dhënave, por analiza kryhet bazuar vetëm në ngjashmërinë e të dhënave me njëra-tjetrën, atëherë parashtrohen disa pyetje të rëndësishme në lidhje metodologjinë e zbatimit të algoritmeve të grupimit: Cili është algoritmi më i përshtatshëm për t'u vënë në punë (duke marrë në konsideratë shkallën e saktësisë së ofruar mesatarisht, shkallën e kompleksitetit që e karakterizon etj) ? Cilat janë parametrat optimalë për sintonizimin e algoritmit? Si do të vlerësohen rezultatet e gjeneruara nga zbatimi i këtyre algoritmeve mbi bashkësitë e të dhënave?

Në këtë kapitull do të zhvillojmë një sërë procedurash eksperimentale për studimin e aplikuar të algoritmeve të grupimit me synimin për t'iu përgjigjur (në mënyrë të plotë ose në mënyrë të pjesshme) pyetjeve të parashtruara më lart. Për të realizuar procedurat eksperimentale kemi shfrytëzuar disa bashkësi të dhënash standarde, disa bashkësi të dhënash sintetike si dhe një bashkësi të dhënash jo-publike. Me termin bashkësi të dhënash standarde i referohemi bashkësive të të dhënave nga bota reale të cilat janë të mirënjohura nga komuniteti i kërkimit shkencor dhe janë të disponueshme publikisht në depozitën e të dhënave të Universitetit të Kalifornisë në Irvine (UCI) [44]. Këto bashkësi janë një referencë mjaft e mirë për studimin e algoritmeve grupues, duke qenë se për shumicën e tyre struktura e brendshme e të dhënave është e njohur. Gjatë procedurave eksperimentale shfrytëzohen vetëm të dhënat dhe numri i grupimeve (pa iu referuar strukturës së brendshme të të dhënave, pra përmbajtjes së grupimeve) dhe vetëm në fund rezulatet e gjeneruara nga zbatimi i algoritmeve të grupimit i krahasojmë me grupimet reale në të dhëna për të përllogaritur indekse të rëndësishëm si shkalla e saktësisë. Bashkësitë e të dhënave sintetike përgjithësisht janë bashkësi entitetesh imagjinare që është vështirë të interpretohen në kornizën e botës reale, të cilat përgjithësisht janë të gjeneruara në mënyrë të automatizuar.

Më tej gjatë këtij kapitulli do të trajtojmë problemin e verifikimit të grupimeve. Vërtet zbatimi i algoritmeve të grupimit mbi bashkësi të dhënash standarde shërben mjaft mirë për qëllime krahasuese eksperimentale, por në thelb qëllimi i algoritmeve të grupimit është eksplorimi i bashkësive të të dhënave për të cilat struktura e brendshme nuk është e njohur. Algoritmet e grupimit gjithnjë gjenerojnë një rezultat ndaj na nevojiten teknika vlerësuese për cilësinë e grupimeve të gjeneruara (pra sa i

përafrohen ato strukturës reale të të dhënave). Indekset e verifikimit të grupimeve janë një mjet mjaft i mirë vlerësues, ndonësë bazohen mbi parime të përgjithshme si kompaktësia e brenda grupimeve, largësia e grupimeve nga njëri-tjetri etj. Në seksionet në vijim do të përshkruajmë procedura eksperimentale që kemi kryer mbi bashkësi të dhënash standarde dhe sintetike, ku grupimet e gjeneruara do t'i vlerësojmë me anë të indekseve të verifikimit.
Në pjesën e fundit të këtij kapitulli do të paraqesim një studim që kemi kryer mbi përdorimin e teknikave të grupimit të turbullt për të eksploruar disa bashkësi të dhënash që u përkasin studentëve të programit të studimit Teknologji Informacioni në Universitetin e Elbasanit. Me anë të këtyre teknikave eksploruese kemi realizuar karakterizimin e studentëve sipas performancës dhe përfundimet e këtij studimi u janë ofruar departamentit për qëllime këshilluese/udhëzuese në lidhje me studentët apo dhe për vendime strategjike në lidhje me menaxhimin e departamentit.

4.1 Bashkësitë e të dhënave të përdorura në eksperimente

Gjatë procedurave eksperimentale të këtij kapitulli kemi shfrytëzuar edhe disa bashkësi sintetike, edhe disa bashkësi standarde të cilat janë të mirënjohura nga komuniteti i kërkimit shkencor dhe janë të disponueshme publikisht në depozitën e të dhënave të Universitetit të Kalifornisë në Irvine [44]. Bashkësitë në fjalë janë Iris, Wine, Glass, Dermatology, Breast Cancer Wisconsin dhe Yeast. Për një përshkrim më të hollësishëm të bashkësive të të dhënave apo për të parë të gjithë elementët e këtyre bashkësive mund të vizitoni faqen e depozitës së të dhënave të cituar, ndërsa në vijim po japim një përshkrim të përmbledhur të këtyre bashkësive të të dhënave, ku përmendim fushën e studimit nga e cila këto të dhëna janë mbledhur, numrin e instancave (pra të elementëve), numrin e atributeve që karakterizon instancat dhe numrin e grupimeve:

- Iris: Kjo bashkësi të dhënash përshkruan karakterisikat e disa luleve të llojit iris. Përmban gjithsej 150 instanca me nga katër atribute (gjatësia dhe gjerësia e petaleve, si dhe gjatësia dhe gjerësia e sepaleve). Intancat janë të shpërndara në tre grupime (Iris Setosa, Iris Versicolour dhe Iris Virginica), por dy nga grupimet kanë mbivendosje të konsiderueshme, ndaj dhe në probleme të analizës së grupimeve për bashkësinë Iris është e pranueshme ndarja si në dy, ashtu dhe në tre grupime.
- Wine: Kjo bashkësi të dhënash përmban rezultatet e analizës kimike të verërave nga i njëjti rajon i Italisë, por nga kultivues të ndryshëm. Gjithsej ndodhen 178 instanca me 13 atribute (të gjitha numerike si: përmbajtja e alkoolit, acidit malik, magnezit etj.) Elementët e bashkësisë janë të shpërndara në gjashtë grupime.

- Glass : Kjo bashkësi të dhënash përshkruan karakteristika fizike dhe kimike të matura eksperimentalisht për disa xhama. Përmban gjithsej 214 instanca me nga nëntë atribute numerike (indeksi i përthyerjes së dritës, përmbajtja e natriumit, magnezit, aluminit, silikonit etj). Për hir të saktësisë përmendim se bashkësia origjinale Glass ka dhjetë attribute numerike por njëri prej tyre është numri rendor dhe natyrisht ky atribut nuk përfshihet në problemin e analizës së grupimeve. Elementët e bashkësisë janë të shpërndara në gjashtë grupime.
- Dermatology: Kjo bashkësi të dhënash përshkruan karakteristika të disa pacientëve në lëkurën e të cilëve ekzistojnë simptoma të zonave me skuqje dhe fortësim (kjo gjendje mund të shkaktohet nga disa sëmundje). Bashkësia gjithsej përmban 366 instanca, me nga 33 atribute numerike (me atribute klinike si niveli i skuqjes, i fortësisë, i kruajtjes, i kufinjve etj, si dhe me atribute histo-patologjike si niveli i melaninës, ekzocitozës, hiperkeratozës etj). Elementët e bashkësisë janë të shpërndara në gjashtë grupime.
- Breast Cancer Wisconsin: Kjo bashkësi përmban të dhëna klinike të pacienteve të diagnostikuara me kancer të gjirit. Gjithsej ndodhen 699 instanca me nga nëntë atribute numerike (trashësia e masivit, uniformiteti i përmasës dhe formës së qelizave, mitoza etj). Për hir të saktësisë përmendim se bashkësia origjinale "Breast Cancer Wisconsin" ka dhjetë attribute numerike por njëri prej tyre është numri rendor dhe natyrisht ky atribut nuk përfshihet në problemin e analizës së grupimeve.
- Yeast: Kjo bashkësi përmban të dhëna rreth proteinave brenda qelizave të majasë ku atributi i klasës paraqet vendndodhjen brenda qelizës. Bashkësia gjithsej përmban 1484 me nga tetë atribute. Elementët e bashkësisë janë të shpërndara në dhjetë grupime.

Në tabelën vijuese po përmbledhim detajet e bashkësive të përdorura:

Bashkësia	**Nr i atributeve**	**Nr i instancave**
Iris	4	150
Wine	13	178
Glass	9	214
Dermatology	33	366
Breast Cancer	9	699
Yeast	8	1484
Synth1	9	250
Synth2	12	320

Tabelë 4 - Detajet e bashkësive të të dhënave të përdorura

4.2 Procedura eksperimentale sintonizuese dhe krahasuese

Për të krahasuar algoritmet kryesorë të grupimit kemi realizuar një sërë eksperimentesh mbi bashkësitë e të dhënave standarde dhe sintetike të përmendura në seksionin e mëparshëm. Qëllimi kryesor i eksperimenteve të kryera ishte krahasimi i algoritmeve në aspektet e shkallës së saktësisë, numrit të përseritjeve (iteracioneve) dhe kohës së ekzekutimit. Për bashkësitë e të dhënave të përdorura grupimet janë të njohura paraprakisht, por nuk i vemë në punë gjatë ekzekutimit të algoritmeve grupues. Grupimet e njohura përdoren vetëm në fazën përfundimtare për të vlerësuar shkallën e saktësisë. Për algoritmet e turbullt para se të kryejmë vlerësimin e shkallës së saktësisë, grupimet e turbullta i shndërrojmë në grupimve të forta. Metodologjia e përdorur për këtë është që çdo element caktohet të jetë anëtar i plotë i grupimit me qendrën e të cilit është më afër. Të gjithë algoritmet e grupimit përmbajnë disa parametra si koeficienti i turbullsisë, funksioni distancë, shkalla e tolerancës etj. Gjatë këtij studimi eksperimental për secilin algoritëm janë realizuar një seri ekzekutimesh me vlera të ndryshme të parametrave. Duke qenë se algoritmet e grupimit janë të ndjeshëm ndaj përzgjedhjes së vlerave fillestare të matricës particion (të cilat vendosen në mënyrë të rastësishme), për secilat vlera të parametrave secili algoritëm është ekzekutuar 25 herë dhe vlera e shfaqur në tabelat pasqyruese është mesatarja e këtyre ekzekutimeve. Në përllogaritjen e distancës midis elementëve të bashkësisë së të dhënave ekziston mundësia që disa atribute ta kenë bashkësinë e vlerave dallueshëm më të madhe se atributet e tjerë dhe si rrjedhim ta dominojnë funksionin distancë. Për të shmangur dukuri të tilla, atributet i normalizojmë paraprakisht. Pra për bashkësinë e të dhënave $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}\}$, ku secili element i saj konceptohet si një vektor shtyllë me d përmasa ku $\boldsymbol{x_i} = [x_{i1}, x_{i2}, .., x_{id}]^T$, atributet normalizohen me anë të $x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$ për çdo $i \in \{1,2, \dots, n\}$ dhe $j \in \{1,2, \dots, d\}$, ku $\bar{x}_j = \frac{\sum_{t=1}^{n} x_{tj}}{n}$ është mesatarja aritmetike dhe $s_j = \sqrt{\left(x_{ij} - \bar{x}_j\right)^2 / (n-1)}$ është devijimi standard i atributit të j-të.

4.2.1 Eksperimente sintonizuese rreth distancave Minkovski

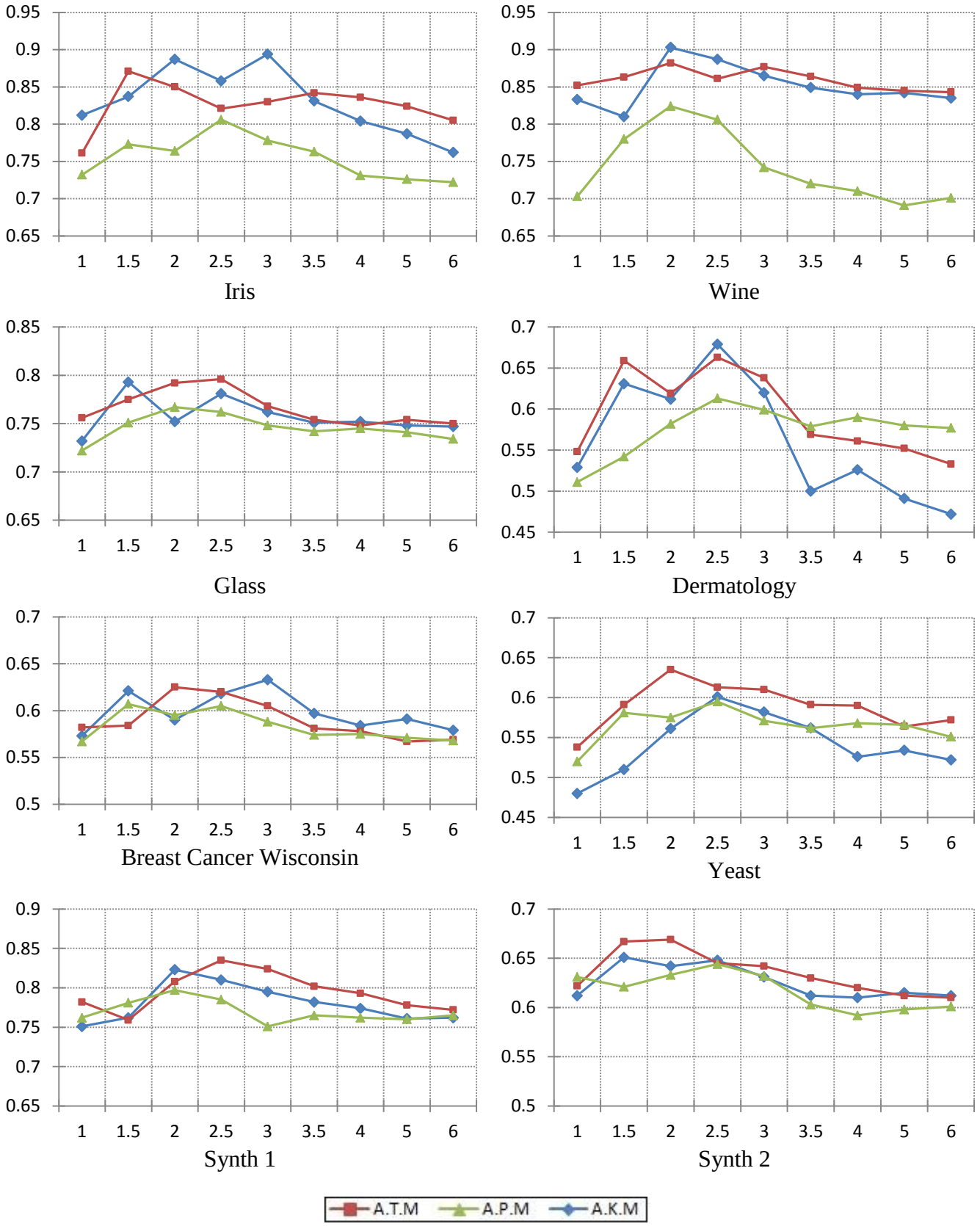
Gjatë trajtimit të çështjes së funksioneve disatancë vumë në dukje se distanca Minkovski është një nga funksionet distancë më të përdorur, madje funksione të tjera distancë (distanca Euklidiane, distanca Manhatan, distanca maksimale) janë raste të veçanta të saj për vlera të ndyshme të rendit (përkatësisht për $m = 2$, $m = 1$ dhe $m \to \infty$). Në këtë seksion do të përshkruajmë disa procedura eksperimentale me karakter sintonizues rreth distancës Minkovski, ku do të variojmë vlerën e rendit të kësaj distance dhe do të aplikojmë algoritmin klasik të mesoreve (A.K.M.), algoritmin

e turbullt të mesoreve (A.T.M.) dhe algoritmin posibilistik të mesoreve (A.P.M.). Kriteri krahasues në këtë eksperiment do të jetë shkalla e saktësisë. Ky eksperiment përveç karakterit sintonizues në lidhje me distancën Minkovski (analizimi i vlerave të rendit për të cilat algoritmet funksionojnë më mirë), shërben edhe si mjet krahasues ndërmjet tri algoritmeve në fjalë. Theksojmë se në këtë eksperiment nuk mund t'i përfshinim dhe algoritmet e tjera të grupimit (si Gustafson-Kessel, Gath-Geva etj) duke qenë se ato bazohen në funksione distancë specialë që shfrytëzojnë matricat e kovariancës. Gjatë këtij eksperimenti vlerat e rendit të distancës Minkovski që kemi përdorur janë përkatësisht 1, 1.5, 2, 2.5, 3, 3.5, 4, 5 dhe 6. Vlerat përkatëse të shkallës së saktësisë për secilin rast janë përmbledhur në tabelën vijuese:

Bashkësia e të dhënave		**Vlera e rendit të distancës Minkovski (m)**								
		1	**1.5**	**2**	**2.5**	**3**	**3.5**	**4**	**5**	**6**
Iris	A.K.M.	0.812	0.837	0.887	0.858	0.894	0.831	0.804	0.787	0.762
	A.T.M.	0.761	0.871	0.850	0.821	0.830	0.842	0.836	0.824	0.805
	A.P.M	0.732	0.773	0.764	0.806	0.778	0.763	0.731	0.726	0.722
Wine	A.K.M.	0.833	0.810	0.903	0.887	0.865	0.849	0.840	0.842	0.835
	A.T.M.	0.852	0.863	0.882	0.861	0.877	0.864	0.849	0.845	0.843
	A.P.M	0.703	0.78	0.824	0.806	0.742	0.720	0.710	0.691	0.701
Glass	A.K.M.	0.732	0.793	0.752	0.781	0.762	0.751	0.752	0.748	0.747
	A.T.M.	0.756	0.775	0.792	0.796	0.768	0.754	0.748	0.754	0.750
	A.P.M	0.722	0.751	0.767	0.762	0.748	0.742	0.745	0.741	0.734
Dermatology	A.K.M.	0.529	0.631	0.612	0.679	0.620	0.500	0.526	0.491	0.472
	A.T.M.	0.548	0.659	0.619	0.663	0.638	0.569	0.561	0.552	0.533
	A.P.M	0.511	0.542	0.582	0.613	0.599	0.579	0.59	0.58	0.577
Breast Cancer	A.K.M.	0.573	0.621	0.590	0.618	0.633	0.597	0.584	0.591	0.579
	A.T.M.	0.582	0.584	0.625	0.620	0.605	0.581	0.578	0.567	0.569
	A.P.M	0.567	0.607	0.595	0.605	0.588	0.574	0.575	0.571	0.568
Yeast	A.K.M.	0.480	0.510	0.561	0.601	0.582	0.562	0.526	0.534	0.522
	A.T.M.	0.538	0.591	0.635	0.613	0.610	0.591	0.590	0.564	0.572
	A.P.M	0.520	0.581	0.575	0.595	0.571	0.562	0.568	0.566	0.551
Synth 1	A.K.M.	0.751	0.762	0.823	0.810	0.795	0.782	0.774	0.761	0.762
	A.T.M.	0.782	0.759	0.808	0.835	0.824	0.802	0.793	0.778	0.772
	A.P.M	0.762	0.781	0.797	0.785	0.751	0.765	0.762	0.760	0.765
Synth 2	A.K.M.	0.612	0.651	0.642	0.648	0.631	0.612	0.610	0.615	0.612
	A.T.M.	0.622	0.667	0.669	0.645	0.642	0.630	0.620	0.612	0.610
	A.P.M	0.631	0.621	0.633	0.644	0.632	0.603	0.592	0.598	0.601

Tabelë 5 - Shkalla e saktësisë në ekperimentet përkatëse në lidhje me distancën Minkovski

Rezultatet e këtij eksperimenti gjithashtu mund t'i paraqesim vizualisht me anë të grafikëve si në faqen në vijim. Për secilën nga tetë bashkësitë kemi ndërtuar një grafik cili paraqet se si varion shkalla e saktësië së tre algoritmeve (A.K.M, A.T.M dhe A.P.M) kur variojmë vlerën e rendit të distancës Minkovski.



Figurë 24 - Shkalla e saktësisë në varësi të rendit të distancës Minkovski për disa bashkësi të dhënash

Në bazë të eksperimenteve të kryera, vumë re se në përgjithësi algoritmet arrinin shkallën më të mirë të saktësisë për vlera të rendit të distancës Minkovski midis 1.5 dhe 3. Për bashkësi të dhënash të cilave nuk u njihet struktura e brendshme paraprakisht, mund të shfrytëzohen indekset e verifikimit të grupimeve (të cilat do të trajtohen në seksionin 4.3) për sintonizimin e rendit të disancës. Gjithashtu vumë re se kur vlera e rendit të distancës Minkovski rritej, ndryshimi në shkallën e saktësisë së algoritmeve zvogëlohej.

Një vëzhgim tjetër i natyrshëm është se për këto bashkësi të dhënash algoritmi posibilistik i mesoreve (A.P.M) ishte inferior në krahasim me algoritmin klasik të mesoreve (A.K.M) dhe algoritmin e turbullt të mesoreve (A.T.M). Vërtet algoritmi posibilistik i mesoreve është më i avantazhuar në bashkësi të dhënash që përmbajnë vlera të largëta ose vlera me zhurmë, por bashkësitë mbi të cilat u kryen eksperimentet nuk karakterizoheshin nga vlera të tilla.

Për sa i përket krahasimit midis algoritmit klasik të mesoreve (A.K.M) dhe algoritmit të turbullt të mesoreve (A.T.M) vëzhguam se për bashkësi të dhënash jo-komplekse (për nga numri i instancave, atributeve dhe grupimeve, siç mund të konsiderohen bashkësitë Iris, Wine, Glass, Breast Cancer Wisconsin dhe Synth1) dallimi në shkallën e saktësisë midis dy algoritmeve ishte mjaft i vogël, ndërsa për bashkësi të dhënash komplekse (siç mund të konsiderohen Dermatology, Yeast dhe Synth2) vihet rë një shkallë pak më e lartë saktësie e algoritmit të turbullt të mesoreve në krahasim me algoritmin klasik të mesoreve.

4.2.2 Eksperimente sintonizuese rreth koeficientit të turbullsisë

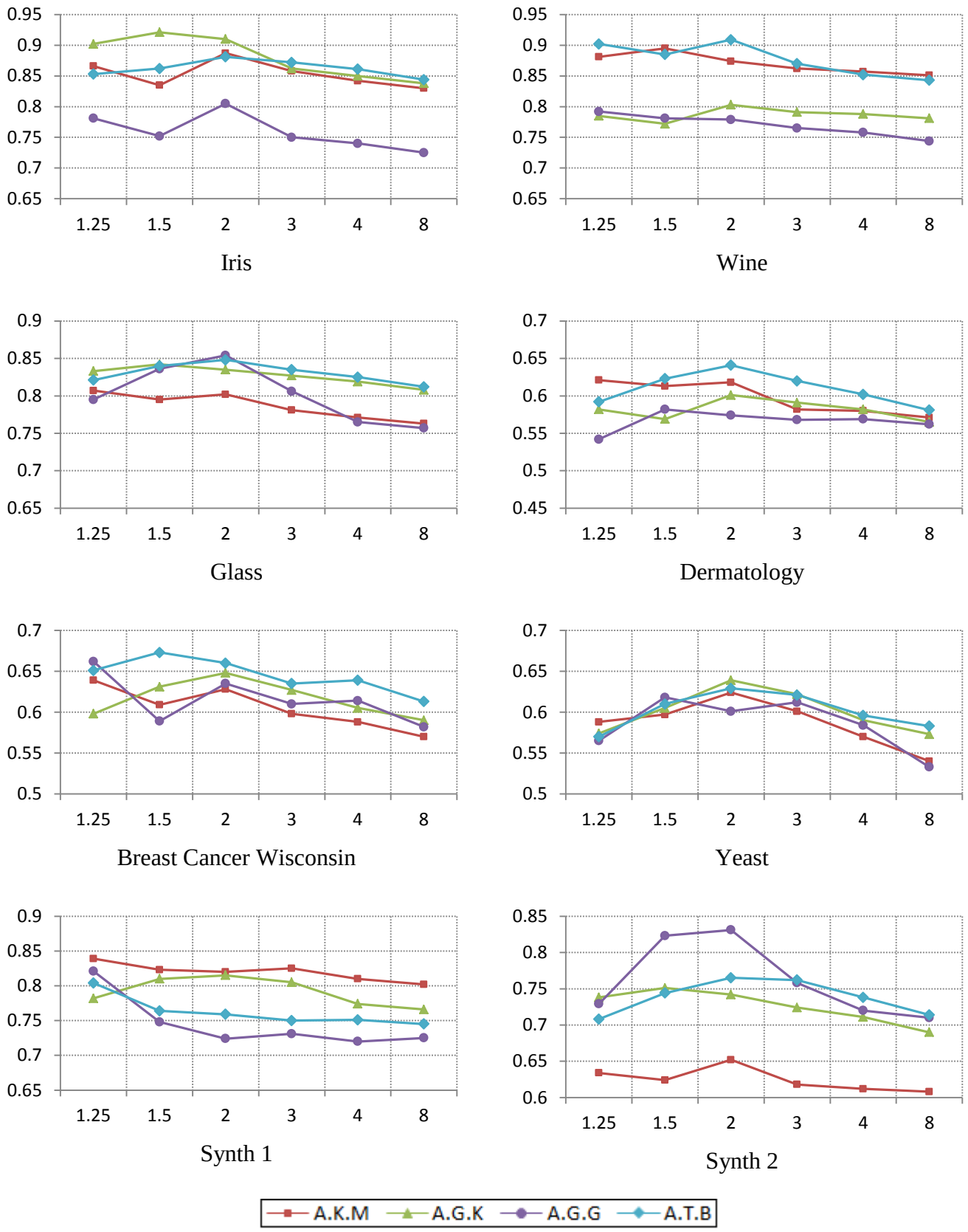
Për algoritmet e turbullta të grupimeve ekzistojnë disa parametra të cilët duhet të specifikohen paraprakisht në mënyrë që algoritmi të fillojë ekzekutimin. Një nga parametrat e rëndësishëm të këtyre algoritmeve është koeficienti i turbullsisë φ. Ky koeficient është një vlerë reale në $[1, +\infty[$, që shpreh se sa të turbullt do të jenë grupimet që do të gjenerohen. Vlera $\varphi = 1$ nënkupton se grupimet janë me kufinj të prerë (klasikë) ndërsa sa më e lartë të jëtë vlera e φ, aq më shumë mbivendosje do të ketë midis grupimeve. Në këtë seksion do të përshkruajmë disa procedura eksperimentale me karakter sintonizues rreth koeficientit të turbullsisë, ku do të variojmë vlerën e këtij koeficienti dhe do të aplikojmë algoritmin e turbullt të mesoreve (A.T.M.), algoritmin Gustafson-Kessel (A.G.K), algoritmin Gath-Geva (A.G.G) dhe algoritmin e turbullt bazuar në bërthama (A.T.B.). Kriteri krahasues në këtë eksperiment do të jetë shkalla e saktësisë. Ky eksperiment përveç karakterit sintonizues në lidhje me koeficientin e turbullsisë (analizimi i vlerave të koefcientit për të cilat algoritmet funksionojnë më mirë), shërben edhe si mjet krahasues

ndërmjet katër algoritmeve në fjalë. Gjatë këtij eksperimenti vlerat e koeficientit të turbullsisë që kemi përdorur janë përkatësisht 1.25, 1.5, 2, 3, 4 dhe 8.

Bashkësia e të dhënave		**Vlera e koeficientit të turbullsisë (φ)**					
		1.25	**1.5**	**2**	**3**	**4**	**8**
Iris	A.T.M.	0.866	0.835	0.887	0.858	0.842	0.830
	A.G.K.	0.902	0.921	0.910	0.862	0.850	0.838
	A.G.G	0.781	0.752	0.805	0.750	0.740	0.725
	A.T.B	0.853	0.862	0.881	0.872	0.861	0.844
Wine	A.T.M.	0.881	0.895	0.874	0.862	0.857	0.851
	A.G.K.	0.785	0.772	0.803	0.791	0.788	0.781
	A.G.G	0.792	0.781	0.779	0.765	0.758	0.744
	A.T.B	0.902	0.885	0.909	0.870	0.852	0.843
Glass	A.T.M.	0.807	0.795	0.802	0.781	0.771	0.763
	A.G.K.	0.833	0.842	0.835	0.827	0.819	0.808
	A.G.G	0.795	0.836	0.854	0.806	0.765	0.757
	A.T.B	0.821	0.840	0.848	0.835	0.825	0.812
Dermatology	A.T.M.	0.621	0.613	0.618	0.582	0.580	0.571
	A.G.K.	0.582	0.569	0.601	0.591	0.582	0.565
	A.G.G	0.542	0.582	0.574	0.568	0.569	0.562
	A.T.B	0.592	0.623	0.641	0.620	0.602	0.581
Breast Cancer	A.T.M.	0.639	0.609	0.628	0.598	0.588	0.570
	A.G.K.	0.598	0.631	0.648	0.627	0.605	0.590
	A.G.G	0.662	0.589	0.635	0.610	0.614	0.582
	A.T.B	0.651	0.673	0.660	0.635	0.639	0.613
Yeast	A.T.M.	0.588	0.597	0.624	0.601	0.570	0.540
	A.G.K.	0.574	0.605	0.639	0.622	0.590	0.573
	A.G.G	0.565	0.618	0.601	0.612	0.584	0.533
	A.T.B	0.570	0.610	0.629	0.621	0.596	0.583
Synth 1	A.T.M.	0.839	0.823	0.820	0.825	0.810	0.802
	A.G.K.	0.782	0.810	0.815	0.805	0.774	0.766
	A.G.G	0.821	0.748	0.724	0.731	0.720	0.725
	A.T.B	0.804	0.764	0.759	0.750	0.751	0.745
Synth 2	A.T.M.	0.634	0.624	0.652	0.618	0.612	0.598
	A.G.K.	0.738	0.751	0.742	0.724	0.711	0.690
	A.G.G	0.729	0.823	0.831	0.758	0.720	0.710
	A.T.B	0.708	0.744	0.765	0.762	0.738	0.714

Tabelë 6 - Shkalla e saktësisë në ekperimentet përkatëse në lidhje me koeficientin e turbullsisë

Rezultatet e këtij eksperimenti gjithashtu mund t'i paraqesim vizualisht me anë të grafikëve si në vijim. Për secilën nga tetë bashkësitë kemi ndërtuar një grafik cili paraqet se si varion shkalla e saktësië së katër algoritmeve (A.T.M, A.G.K, A.G.G dhe A.T.B) kur variojmë vlerën e koeficientit të turbullsisë.



Figurë 25 - Shkalla e saktësisë në varësi të koeficientit të turbullsisë për disa bashkësi të dhënash

Në bazë të eksperimenteve të kryera, vumë re se në përgjithësi algoritmet arrinin shkallën më të mirë të saktësisë për vlera të koeficientit të turbullsisë midis 1.25 dhe 2. Për bashkësi të dhënash të cilave nuk u njihet struktura e brendshme paraprakisht, mund të shfrytëzohen indekset e verifikimit të grupimeve (të cilat do të trajtohen në seksionin 4.3) për sintonizimin e koeficientit të turbullsisë. Gjithashtu vumë re se në të katërt algoritmet e zbatuar kur vlera e koeficientit të turbullsisë rritej, shkalla e saktësisë zvogëlohej, si dhe ndryshimi në shkallën e saktësisë midis algoritmeve të ndryshme zvogëlohej.

Në krahasimin e përgjithshëm të algoritmeve më njëri-tjetrin vëzhgojmë se algoritmi Gath-Geva (A.G.G) karakterizohet nga një luhatshmëri më e madhe në krahasim me algoritmet e tjerë (pra ka luhatjet më të mëdha në shkallën e saktësisë kur varion koeficienti i turbullsisë). Kjo vjen si pasojë e termit eksponencial në shprehjen e distancës për këtë algoritëm që ndikon që ai të konvergjojë më shpejt në optimume lokale. Avantazhi i këtij algoritmi është i dallueshëm në raste kur kemi të bëjmë me bashkësi që përmbajnë grupime të formave dhe dendësive të ndryshme, siç është rasti i bashkësisë Synth2.

Algoritmi Gustafson-Kessel gjithashtu vëzhgohet të ketë një shkallë të mirë saktësie, madje për disa bashkësi të dhënash (si Iris dhe Dermatology) ka rezultuar të jetë më i suksesshmi.

Algoritmi i bazuar në bërthamë është një algoritmëm i qëndrueshëm, i cili shpeshherë është i pari në krahasim me të tjerët, por edhe në rastet kur nuk është kryesues, është nga algoritmet me shkallë saktësie më të lartë.

4.2.3 Eksperimente krahasuese rreth performancës së algoritmeve të grupimit

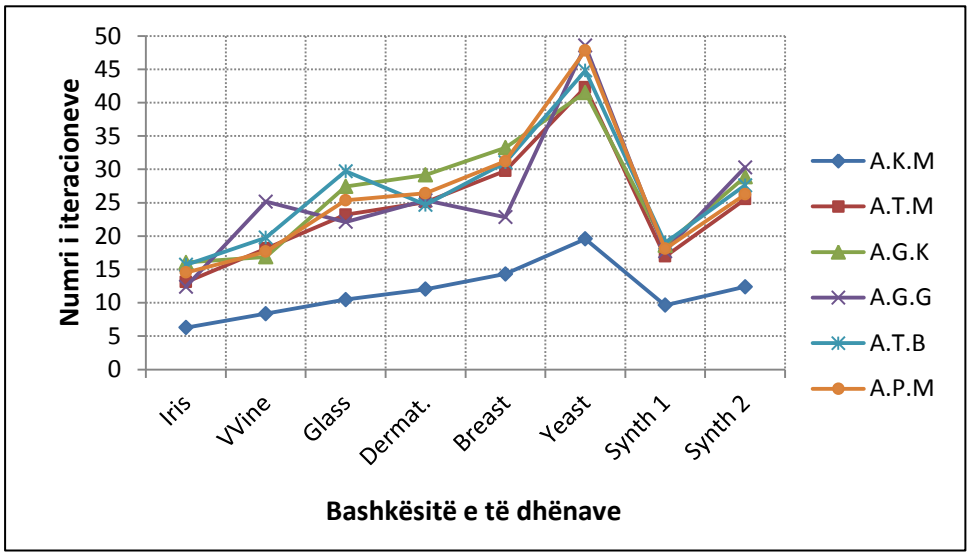
Në këtë seksion do të realizojmë eksperimente krahasuese në lidhje me algoritmet e grupimit nga këndvështrimi i performancës, pra nga numri i përsëritjeve (iteracioneve) dhe koha e ekzekutimit. Theksojmë se krahasimet që po realizojmë në këtë seksion konsiderohen si dytësore, sepse kriteri parësor vlerësues i algoritmeve të grupimit është shkalla e saktësisë. Gjithsesi njohja e performancës së algoritmeve të grupimit në disa raste ka një rëndësi të veçantë si për shembull në projektimin e metodave ansambël (të cilat do t'i hulumtojmë në kapitullin e gjashtë).

Edhe këto eksperimente janë kryer mbi të njëjtat bashkësi të dhënash si dhe eksperimentet e kaluara (pra bashkësitë e përshkruara në seksionin 4.1). Secili algoritëm është ekzekutuar 25 herë dhe vlera e paraqitur në tabela është mesatarja aritmetike e numrit të persëritjeve (iteracioneve) të algoritmit në bashkësinë përkatëse. Për të gjithë algoritmet që përdorin koeficientin e turbullsisë, gjatë këtyre eksperimenteve është përdorur vlera $\varphi = 2$, si dhe për algoritmet A.K.M, A.T.M dhe A.P.M si funksion distancë është përdorur distanca Euklidiane.

Bashkësia e të dhënave	Numri mesatar i përsëritjeve (iteracioneve)					
	A.K.M	A.T.M	A.G.K	A.G.G	A.T.B	A.P.M
Iris	6.28	13.08	16.04	12.40	15.68	14.52
Wine	8.36	18.12	16.88	25.16	19.76	17.64
Glass	10.48	23.20	27.44	22.12	29.72	25.34
Dermatology	12.04	25.08	29.16	25.36	24.68	26.40
Breast Cancer	14.32	29.76	33.24	22.84	31.04	31.20
Yeast	19.56	42.28	41.52	48.56	44.80	47.80
Synth 1	9.64	16.92	18.60	17.72	19.00	18.08
Synth 2	12.40	25.52	28.88	30.28	27.60	26.24

Tabelë 7- Numri mesatar i përsëritjeve (iteracioneve) për algoritmet përkatëse

Të dhënat e tabelës gjithashtu mund të paraqiten në mënyrë grafike si në figurën 26:



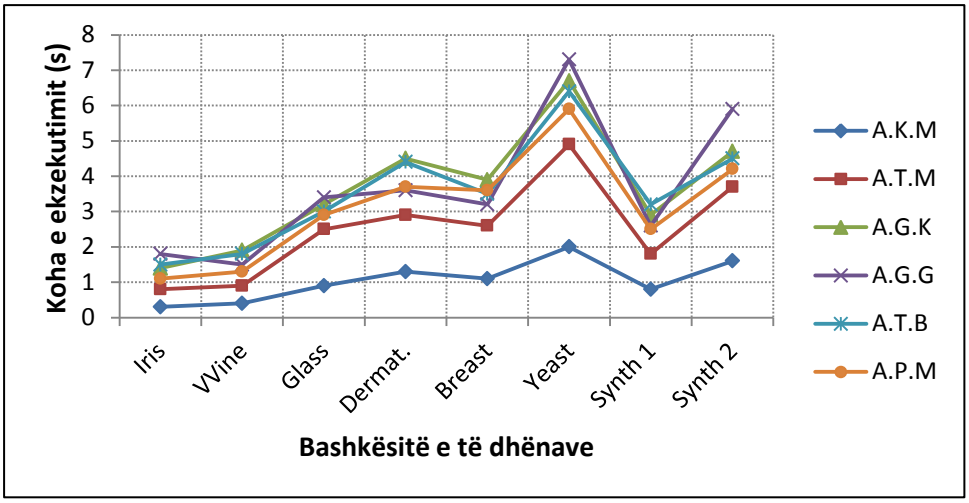
Figurë 26 - Numri i iteracioneve të algoritmeve për disa bashkësi të dhënash

Për të njëjtat eksperimente të pasqyruara në tabelën dhe grafikun e mësipërm, përveç numrit të iteracioneve, kemi matur dhe kohën e ekzekutimit të tyre dhe mesataret përkatëse i paraqesim në tabelën në vijim.

Bashkësia e të dhënave	Koha mesatare e ekzekutimit (në sekonda)					
	A.K.M	A.T.M	A.G.K	A.G.G	A.T.B	A.P.M
Iris	0.3	0.8	1.4	1.8	1.5	1.1
Wine	0.4	0.9	1.9	1.5	1.8	1.3
Glass	0.9	2.5	3.2	3.4	3.0	2.9
Dermatology	1.3	2.9	4.5	3.6	4.4	3.7
Breast Cancer	1.1	2.6	3.9	3.2	3.5	3.6
Yeast	2.0	4.9	6.7	7.3	6.4	5.9
Synth 1	0.8	1.8	2.9	2.6	3.2	2.5
Synth 2	1.6	3.7	4.7	5.9	4.5	4.2

Tabelë 8 - Koha mesatare e ekzekutimit (në sekonda) për algoritmet përkatëse

Të dhënat e tabelës gjithashtu mund të paraqiten në mënyrë grafike si në figurën 27:



Figurë 27 - Koha e ekzekutimit të algoritmeve për disa bashkësi të dhënash

Nga grafiku i numrit të iteracioneve vëzhgimi kryesor është se algoritmi klasik i mesoreve (A.K.M) përfundon me një numër më të vogël iteracionesh në krahasim me të gjithë algoritmet e tjerë. Nga ana tjetër, për sa i përket numri të iteracioneve nuk ka ndonjë dallim të konsiderushëm midis algoritmeve të tjerë.

Më tej nga grafiku i kohës së ekzekutimit, vëzhgojmë se algoritmi klasik i mesoreve (A.K.M) ka një kohë ekzekutimi më të vogël, dallueshëm nga të gjithë algoritmet e tjerë. Gjithashtu ne grupin e algoritmeve të tjerë (pra përveç A.K.M), për sa i përket kohës së ekzekutimit algoritmi i turbullt i mesoreve (A.T.M) ka një kohë ekzekutimi më të vogël në krahasim me të tjerët (A.G.K, A.G.G, A.T.B dhe A.P.M). Arsyeja që kemi dallim në kohën e ekzekutimit ndërkohë që nuk kishim ndonjë dallim të konsiderueshëm në numrin e iteracioneve, është se iteracionet e algoritmeve të tjera janë më të "kushtueshme" se iteracionet e A.T.M. Për shembull gjatë një iteracioni të algoritmit Gustafson-Kessel apo Gath-Geva duhet të llogaritet matrica e kovariancës si dhe e anasjellta e saj. Në rast se do të kishim të bënim me bashkësi të dhënash me numër akoma më të madh elementësh, apo me numër akoma më të madh atributesh për secilin element, dallimi ndërmjet A.T.M dhe algoritmeve të tjerë për sa i përket kohës së ekzekutimit do të pritej të ishte akoma më i madh.

4.3 Verifikimi i grupimeve

Një problem mjaft i rëndësishëm i cili synon të plotësojë procedurën e analizës së grupimeve është verifikimi i grupimeve. Në morinë e algoritmeve të grupimit mund të përzgjedhim cilindo për ta zbatuar mbi një bashkësi të dhënash, por çështja thelbësore është të vlerësohet se sa pranë strukturës së vërtetë të bashkësisë së të dhënave është

rezultati i gjeneruar. Në shembujt e seksionit të mëparshëm struktura e brendshme e të dhënave ishte e njohur paraprakisht dhe mund ta shfrytëzonim në mënyrë të drejtpërdrejtë për të vlerësuar grupimet e gjeneruara. Kjo ishte mjaft e dobishme për qëllime eksperimentale, por në thelb qëllimi i algoritmeve të grupimit është eksplorimi i bashkësive të të dhënave për të cilat nuk është e njohur paraprakisht struktura e brendshme.

Gjithashtu në procedurat eksperimentale të seksionit të kaluar ne vërtet nuk shfrytëzuam informacionin e disponueshëm mbi grupimet ekzistuese gjatë ekzekutimit të algoritmit grupues, por ne shfrytëzuam të gatshëm numrin e grupimeve (i cili është një nga parametrat e nevojshëm për t'u specifikuar praprakisht për algoritmet e grupimit). Në bashkësi të dhënash për të cilat nuk është e njohur paraprakisht struktura e brendshme, një nga problemet më sfidues është përcaktimi i numrit optimal të grupimeve. Pra mund të përmbledhim duke vemë në dukje tri qëllime kryesore të trajtimit të problemit të verifikimit të grupimeve:

- Krahasimi rezultateve të gjeneruar nga algoritme të ndryshëm (ose nga i njëjti algoritëm i ekzekutuar me vlera të ndryshme të parametrave)
- Përcaktimi i numrit të grupimeve
- Shmangia e gjetjes së grupimeve në të dhëna me zhurmë

Për trajtimin e këtyre problemeve është zhvilluar një metodologji e cila shfrytëzon disa indekse verifikues të cilët bazohen në dy karakteristika thelbësore: kompaktësia dhe ndarja e grupimeve. Kompaktësia është një madhësi që vlerëson variacionin e të dhënave brenda grupimeve ndërsa ndarja është një madhësi që përshkruan distancat midis grupimeve. Qëllimi i të gjitha teknikave të verifikimit të grupimeve është të zvogëlojnë kompaktësinë dhe të rrisin ndarjen [45]. Gjatë studimeve eksperimentale ne do të jemi veçanërisht të fokusuar në trajtimin e problemit të përcaktimit të numrit optimal të grupimeve në një bashkësi të dhënash. Dy përqasjet kryesore ndaj këtij problemi janë:

- Përzgjedhim një ose disa indekse vlerësues (të cilët do t'i trajtojmë në seksionin vijues). Përcaktohet një vlerë maksimale e arsyeshme për numrin maksimal të grupimeve që mund të ketë bashkësia e të dhënave në fjalë (le të jetë kjo vlerë k_{max}). Më pas për çdo numër natyror k nga 2 në k_{max} ekzekutojmë algoritmin grupues ku vlerë e parametrit të numrit të grupimeve të jetë k. Secilën ndarje në grupime të përftuar e vlerësojmë më anë të indeksit/indekseve të verifikimit dhe në fund përzgjedhim vlerën e k-së e cila e optimizon indeksin/indekset verifikues.
- Në përqasjen e dytë supozojmë se kemi një indeks verifikues për vlerësimin e secilit grupim veçmas. Përcaktohet një vlerë maksimale e arsyeshme për numrin maksimal të grupimeve që mund të ketë bashkësia e të dhënave në fjalë (le të jetë kjo vlerë k_{max}). Algoritmi grupues ekzekutohet për vlerën k_{max} të parametrit të numri të grupimeve. Grupimet e përftuara i vlerësojmë

në bazë indeksit verifikues. Grupimet e ngjashme shkrihen në grupime më të mëdha në mënyrë që të mos ketë grupime "jo-të përshtatshme". Kjo procedurë përsëritet deri në optimizimin e ndarjes në grupime.

Siç mund të kuptohet lehtë, problemi i verifikimit të grupimeve është një problem sfidues me një shkallë kompleksiteti njehsues mjaft të lartë.

4.3.1 Indekset e verifikimit

Në këtë seksion do të shpjegojmë konkretisht disa nga indekset e verifikimit të përdorur më gjerësisht. Është i këshillueshëm zbatimi i disa indekseve verifikues njëkohësisht për të realizuar në mëmyrë sa më të saktë verifikimin e grupimeve, duke qenë se një indeks nuk ka kapacitetin për të qenë vendimtar për të gjitha rastet e ndarjes në grupime [46]. Në shprehjet matematikore në vijim kemi ruajtur të njëjtat konvencione të shënimeve si në kapitujt e mëparshëm, pra k është numri i grupimeve, n është numri i elementëve në bashkësinë e të dhënave, c_i shpreh qendrën e grupimit të i-të dhe $\mu_{i,j}$ shpreh vlerën e anëtarësisë së elementit të j-të në grupimin e i-të.

- *Koeficienti i particionit* (shënohet PC – partition coefficient) shpreh sasinë e mbivendosjes midis grupimeve dhe përllogaritet si:

$$PC(k) = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mu_{i,j}\right)^2$$

  Vlerat e këtij indeksi plotësojnë mosbarazimin $\frac{1}{k} \leq PC(k) \leq 1$. Disavantazh i këtij indeksi është mungesa e lidhjes së drejtpërdrejtë me ndonjë veti të vetë të dhënave si dhe zvogëlohet monotonisht në lidhje me k. Vlera optimale e k-së është ajo që e maksimizon vlerën e këtij indeksi [47].
- *Entropia e particionit* (shënohet PE – partition entropy) shpreh sasinë e turbullsisë në grupime dhe përllogaritet si:

$$PE(k) = -\frac{1}{N}\sum_{i=1}^{c}\sum_{j=1}^{N}\mu_{i,j}\ln{\mu_{i,j}}^2$$

  Për ndarjet probabilistike Bezdek ka vërtetuar se $0 \leq 1 - PC(k) \leq PE(k)$. Vlera optimale e k-së është ajo që e minimizon vlerën e këtij indeksi.
- *Koeficienti i particionit i modifikuar* (shënohet MPC – modified partition coefficient) përllogaritet si:

$$MPC(k) = 1 - \frac{k}{k-1}\left(1 - PC(k)\right)$$

  Si koeficienti i particionit, ashtu dhe entropia e particionit kanë prirje variacioni monoton në lidhje me vlerën e k-së, ndërsa ky indeks e zvogëlon

këtë prirje. Indeksi plotëson mosbarazimin $0 \leq MPC(k) \leq 1$. Vlera optimale e k-së është ajo që e maksimizon vlerën e këtij indeksi [46].

- *Indeksi i ndarjes* (shënohet SI – separation index) shpreh raportin e kompaktësisë totale me vlerën më të vogël të distancës ndarjëse midis grupimeve. Përllogaritet si:

$$SI(k) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mu_{i,j}\right)^{\varphi}\left\|x_j - c_i\right\|^2}{n \cdot \min\limits_{i,j}\left\|c_j - c_i\right\|^2}$$

Ky indeks është i përshtatshëm për t'u përdorur në krahasimin e particioneve që përbëhen nga i njëjti numër grupimesh të turbullta. Vlera optimale e k-së është ajo që e maksimizon vlerën e këtij indeksi [46],[47].

- *Indeksi Xie-Beni* (shënohet XB) përllogaritet si:

$$XB(k) = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mu_{i,j}\right)^{\varphi}\left\|x_j - c_i\right\|^2}{n \cdot \min\limits_{i,j}\left\|x_j - c_i\right\|^2}$$

Në pamje të parë ky indeks duket pothuaj i njëjtë me indeksin e ndarjes, por praktikisht ndryshimi i "vogël" në emëruesin e shprehjes sjell një dallim të konsiderueshëm. Vlera optimale e k-së është ajo që e minimizon vlerën e këtij indeksi [46],[47].

- Indeksi Fukuyama-Sugeno (shënohet FS) përllogaritet si:

$$FS(k) = \sum_{i=1}^{k}\sum_{j=1}^{n}\left(\mu_{ij}\right)^{\varphi}\left(x_j - c_i\right)^2 - \sum_{i=1}^{c}\sum_{j=1}^{n}\left(\mu_{ij}\right)^{\varphi}\left(c_i - \bar{c}\right)^2$$

ku $\bar{c} = \left(\sum_{i=1}^{k} c_i\right)/k$ (mesatarja e qendrave të grupimeve). Vlera optimale e k-së është ajo që e maksimizon vlerën e këtij indeksi [46].

- Indeksi i Dunn (shënohet DI) përllogaritet si:

$$DI(k) = \min_{i \in c}\left\{\min_{j \in c, i \neq j}\left\{\frac{\min\limits_{x \in C_i y \in C_j,} d(x,y)}{\max\limits_{k \in C}\{\max\limits_{x,y \in C} d(x,y)\}}\right\}\right\}$$

Fillimisht ky indeks është projektuar për të dalluar grupime të ndara qartësisht. Disavantazhi kryesor i këtij indeksi është se me rritjen e numrit të elementëve dhe të numrit të grupimve, rritet ndjeshëm kompleksiteti njehsues për këtë indeks. Vlera optimale e k-së është ajo që e maksimizon vlerën e këtij indeksi.

- Indeksi alternativ i Dunn (shënohet ADI) përllogaritet si:

$$ADI(k) = \min_{i \in c}\left\{\min_{j \in c, i \neq j}\left\{\frac{\min\limits_{x \in C_i y \in C_j,} \left|d\left(y, v_j\right) - d\left(x_i, v_j\right)\right|}{\max\limits_{k \in C}\{\max\limits_{x,y \in C} d(x,y)\}}\right\}\right\}$$

Siç shihet, ky indeks është një variacion i indeksit origjinal të Dunn me qëllim parësor përmirësimin e komplkesitetit njehsues. Në vend të përdorimit të drejtpërdrejtë të distancës $d(x, y)$, ne përdorim një madhësi më të vogël si $|d(y, v_j) - d(x_i, v_j)|$ Është e qartë së madhësia e dytë është më e vogël se e para nga mobarazimi i trekëndëshit.

- Hipervolumet e turbullta (shënohet FHV – fuzzy hypervolumes) konceptohet si vëllimi i grupimit të turbullt dhe dhe përllogaritet si:

$$\mathrm{FHV}(k) = \sum_{i=1}^{k} \det(F_i)$$

Pra shprehet shumë e përcaktorëtve të matricave [46]:

$$F_i = \frac{\sum_{j=1}^{n} \mu_{ij}^{\varphi} (x_j - c_i)(x_j - c_i)^T}{\sum_{j=1}^{N} \mu_{ij}^{\varphi}}$$

4.3.2 Rezultate ekperimentale mbi indekset e verifikimit

Indekset e verifikimit të përmendura në seksionin e mëparshëm i kemi implementuar mbi dy bashkësi të dhënash sintetike (Synth1 dhe Synth2) për të trajtuar problemin e gjetjes së numrit optimal të grupimeve në to duke zbatuar algoritmin e turbullt të mesoreve. Përqasja që kemi zbatuar në këto ekperimente ishte ekzekutimi i algoritmit disa herë mbi secilën bashkësi për vlera të numrit të grupimeve që variojnë nga 2 në 10 (një vlerë që e gjykojmë si mjaftueshëm të madhe). Grupimet e gjeneruara nga secili ekzekutim i vlerësojmë me anë të indekseve të grupimit. Rezultatet janë pasqyruar në tabelën vijuese:

k	$PC(k)$	$PE(k)$	$MPC(k)$	$SI(k)$	$XB(k)$	$FS(k)$	$DI(k)$	$ADI(k)$	$FHV(k)$
2	0.72845	0.23655	0.60872	0.0154	21.5271	0.2201	0.1131	0.0024	0.8474
3	0.68715	0.3184	0.59685	0.0169	7.7585	0.8954	0.0311	0.0019	1.0251
4	0.6775	0.3394	0.60428	0.0173	5.3847	0.6210	0.0188	0.0012	1.9141
5	0.6711	0.3655	0.59977	0.0127	6.2213	0.0782	0.0225	0.0008	2.2108
6	0.66455	0.39395	0.59746	0.0104	7.8739	-0.5841	0.0265	0.0006	2.0034
7	0.65695	0.41275	0.58887	0.0099	9.3617	-1.2018	0.0179	0.0005	1.7105
8	0.65375	0.4223	0.57001	0.0094	11.1568	-1.8841	0.0199	0.0001	1.3912
9	0.64165	0.4541	0.53072	0.0102	14.7286	-2.6810	0.0270	0.0002	1.0154
10	0.64785	0.4510	0.4569	0.0092	16.8604	-3.7541	0.0192	0.0003	0.9454

Tabelë 9 - Rezultate eksperimentale të indekseve të verifikimit mbi bashkësinë Synth 1

I njëjti eksperiment u krye gjithashtu mbi bashkësinë të dhënave Synth2 dhe rezultatet janë pasqyruar në tableën vijuese:

k	$PC(k)$	$PE(k)$	$MPC(k)$	$SI(k)$	$XB(k)$	$FS(k)$	$DI(k)$	$ADI(k)$	$FHV(k)$
2	0.9713	0.3171	0.8725	0.0161	11.1959	0.5214	0.1304	0.00594	1.5153
3	0.9162	0.5905	0.8743	0.0197	12.7646	0.7412	0.0358	0.00226	2.3875
4	0.9033	0.6692	0.8711	0.0170	14.4532	0.8954	0.0217	0.00178	2.9123
5	0.8948	0.7721	0.8685	0.0104	8.7432	0.6542	0.0260	0.00096	2.7997
6	0.8861	0.8904	0.8633	0.0094	10.9261	0.2879	0.0305	0.00081	2.7034
7	0.8759	0.9721	0.8553	0.0106	12.4278	-0.4219	0.0207	0.00064	2.6895
8	0.8717	1.0147	0.8533	0.0101	14.8412	-1.3510	0.0229	0.00052	2.2985
9	0.8555	1.1618	0.8375	0.0109	15.206	-3.0854	0.0312	0.00044	1.7730
10	0.8638	1.1471	0.8487	0.0099	15.5387	-4.1576	0.0221	0.00030	1.6674

Tabelë 10 - Rezultate eksperimentale të indekseve të verifikimit mbi bashkësinë Synth 2

Numrat optimalë të grupimeve sipas secilit indeks i përmbledhim në tabelën vijuese:

Indeksi i verifikimit	Synth1	Synth2
$PC(k)$	2	2
$PE(k)$	2	2
$MPC(k)$	2	3
$PI(k)$	3	4
$SI(k)$	4	2
$XB(k)$	4	5
$FS(k)$	3	4
$DI(k)$	2	2
$ADI(k)$	2	2
$FHV(k)$	5	3

Tabelë 11 - Përmbledhje e procedurës verifikuese me numrin optimal të grupimeve për secilin rast

Në bazë të rezultateve eksperimentale të paraqitura në tabelat përmbledhësë të mësipërme theksohet akoma më shumë ideja se procesi i verifikimit të grupimeve është mjaft delikat. Zbatimi i vetëm një indeksi të verifikimit nuk jep një garanci të mjaftueshme për një vlerësim të qëndrueshëm të cilësisë së grupimeve. Nga ana tjetër zbatimi i disa indekseve të verifikimit ofron një qëndrueshmëri më të mirë për një vlerësim sa më objektiv të grupimeve të përftuara nga algoritmet e grupimit. Në shembullin tonë për bashkësitë Synth1 dhe Synth2 situata është mjaft e ngjashme. Në bazë të rezultateve të pasqyruara mund të pranohen dy ose tre grupime (madje, ndonëse është më pak e mundshme, për numrin e grupimeve të këtyre bashkësive nuk mund të kundërshtohet as vlera katër). Kjo vjen si pasojë e pranisë së grupimeve midis të cilave ka shumë mbivendosje ndaj për to ekziston mundësia të konceptohen si një grupim i vetëm). Pavarësisht se struktura e këtyre bashkësive të të dhënave ishte e tillë që mund të sugjeroheshin disa vlera (dhe jo një vlerë e vetme) për numrin optimal të grupimeve, përsëri ky rezultat konsiderohet mjaft informues, pasi në fillim numri optimal i grupimeve ishtë një madhësi plotësisht e panjohur.

4.4 Analiza e grupimeve në eksplorimin e të dhënave të një sistemi arsimor

Gjatë viteve të fundit dixhitalizimi i të dhënave të detajuara të studentëve në universitetet shqiptare jo vetëm ka thjeshtëzuar procedurat tipike të menaxhimit të universitetit, por gjithashtu ka krijuar mundësi të reja për realizimin e një analize të thelluar të të dhënave. Zbatimi i teknikave të analizës së grupimeve mbi këto bashkësi të dhënash mundëson zbulimin e strukturave të brendshme të “fshehura” në të dhëna të cilat mund të shfrytëzohen në shumë çështje strategjike si krijimi i profileve karakterizues të performancës së studentëve, përmirësimi i këshillimit/rekomandimit për studentët, proceset e vendimmarrjes etj.

Në studimin tonë kemi vënë në përdorim (në mënyrë të autorizuar) të dhëna të sistemit të informacionit në Universitetin e Elbasanit në lidhje me studentët që janë të regjistruar në programin e Teknologjisë së Informacionit. Mbi këto bashkësi të dhënash kemi ekzekutuar një sërë algoritmesh të grupimit të turbullt si algoritmi i turbullt i mesoreve, algoritmi Gustafson-Kessel, algoritmi i turbullt bazuar në bërthama etj. Ndër rezultatet e gjeneruara kemi përzgjedhur grupimet më cilësore duke u bazuar në indekset e verifikimit të grupimit. Më tej rezultatet e përzgjedhura janë interpretuar dhe shfrytëzuar në dy drejtime kryesore:

- krijimin e profileve karakterizues të përgjithshëm të performancës së studentëve (që ndihmojnë në procese analizuese dhe këshilluese)
- krijimin e profileve karakterizues të performancës në gruplëndë të caktuara (të cilat ndihmojnë veçanërisht në rekomandimin e lëndëve me zgjedhje)

4.4.1 Metodologjia e zbatuar

Siç përmendëm, ideja qëndrore e këtij studimi ishte zbatimi i disa algoritmeve të grupimit të turbullt në bashkësitë e të dhënave të studentëve për të zbuluar grupime që përfaqësjonë profile karakterizues të performancës së studentëve. Elementët e bashkësive në fjalë përmbajnë një numër të madh elementësh dhe atributesh. Analiza e grupimeve në hapësirat me shumë atribute është një detyrë sfiduese, jo vetëm se është një proces me intesitet të lartë përllogaritës, por gjithashtu prezenca e vlerave të largëta dhe e vlerave me zhurmë ndikon ndjeshëm në rezultatet e gjeneruara [48]. Në rastin tonë elementët përmbajnë të dhëna të disa llojeve të ndryshme si diskrete, të vazhdueshme apo kategorike. Përfshirja e të gjithë këtyre atributeve njëkohësisht në procesin e analizës së grupimeve do të shoqërohej me vështirësi në implementim dhe shkallë mjaft të lartë të kompleksitetit njehsues. Në këto rrethana hapi i parë përpunues i bashkësisë së të dhënave ishte ulja e numrit të atributeve (që njihet si etapa e reduktimit të përmasës), të cilën e realizuam sipas tri strategjive kryesore: përzgjedhja e atributeve, kuantizimi dhe bashkimi.

Me anë të përzgjedhjes së atributeve, vetëm atributet më të rëndësishëm (më përfaqësues) të elementëve të bashkësive përfshihen në përllogaritjen e distancës midis elementëve. Përzgjedhja mund të kryhet manualisht nga ekspertë të fushës përkatëse, por gjithashtu ekzistojnë dhe teknika të automatizuara. Nga ana tjetër me anë të kuantizimit lloje të ndryshme të dhënash të vazhdueshme transformohen dhe përshtaten në formë diskrete. Përdorimi i kuantizimit mund të shoqërohet me humbje precizioni, por në përgjithësi kjo ndikon pak në rezultatin e algoritmeve. Përfundimisht me anë të metodës së bashkimit (agregimit) ne kombinojmë dhe shkrijmë disa atribute në një atribut të vetëm të ponderuar. Vlera e këtij atributi do të jetë përfaqësuesja e cila do të përfshihet në përllogaritjen e distancës midis elementëve.

Pasi është plotësuar ulja e numrit të atributeve të elementëve në bashkësinë e të dhënave (pra reduktimi i përmasës), zbatojmë nga disa herë algoritmin e turbullt të mesoreve, algoritmin Gustafson-Kessel, algoritmin posibilistik të mesoreve dhe algoritmin e turbullt të mesoreve bazuar në bërthama. Këto algoritme u ekzekutuan me vlerë të parametrit të turbullsisë $\varphi = 2$ dhe me vlera të numri të grupimevë që varionin nga 2 në 10. Më tej të gjithë grupimet e gjeneruara u vlerësuan duke përdorur tri indekse të verifikimit të grupimeve (koeficientin e particionit, indeksin Xie-Beni dhe hipervolumet e turbullta) duke përzgjedhur numrin optimal të grupimeve dhe ndarjen më të mirë në grupime për këtë numër grupimesh. Përfundimisht ndarja optimale në grupime u jepet ekspertëve për të realizuar interpretimin e tyre. Në një stad më të zhvilluar të këtij studimi kemi planifikuar gjithashtu zhvillimin e teknikave më të avancuara për të realizuar në mënyrë të automatizuar etapën përfundimtare (interpretimin e grupimeve të gjeneruara) [48].

4.4.2 Krijimi i profileve karakterizues të performancës së përgjithshme

Në universitete është e zakonshme që stafit akademik u caktohet në mënyrë të përkohshme detyra e mentorit (këshilluesit) të klasave apo grupeve studentësh për t'i ndihmuar ata duke i orientuar në aspekte të ndryshme. Detyra e mentorit konsiderohet e përbërë nga dy pjesë themelore: pjesa e përgjithshme (këshilla që janë të njëjta për grupe/klasa studentësh) si informacion rreth lëndëve të ofruara sipas semestrave, rregullave të përgjithshme të procedurave universitare, shfrytëzimit të burimeve të ofruara nga universiteti etj, si dhe pjesa specifike (këshilla që janë specifike në varësi të studentit) si t'i japë rekomandime për lëndët me zgjedhje apo në përzgjedhjen e profilit (për programe studimi të cilat në vitin e fundit ofrojnë disa profile të ndryshme), t'i këshillojë në lidhje me vështirësitë/performancat e dobëta që ata mund të përballen. Procesi i këshillimit të studentëve është i vështirë për t'u trajtuar në mënyrë specifike [48]. Studimi ynë synon t'u ofrojë mentorëve më shumë informacion të dobishëm rreth profileve të performancës së studentëve duke u përmbledhur atyre rezultatet e analizës së turbullt të grupimeve.

Si një shembull ilustrues po përshkruajmë rezultatet e analizës së turbullt të grupimeve të zbatuar mbi të dhënat e tre semestrave të parë të studentëve të regjistruar në vitin 2011-2012 në programin e studimit Teknologji Informacioni. Duke ndjekur metodologjinë e përshkruar në seksionin e mëparshëm, studimi ynë ka kaluar në disa etapa. Fillimisht kemi realizuar reduktimin e përmasës për elementët e bashkësisë së të dhënave duke përzgjedhur atributet më përfaqësuese. Më pas kemi aplikuar një sërë algoritmesh të turbullt mbi të dhënat si algoritmi i turbullt të mesoreve, algoritmi Gustafson-Kessel, algoritmi posibilistik i mesoreve dhe algoritmi i turbullt i bazuar në bërthama, ku secili algoritëm është ekzekutuar me vlera të numrit të grupimeve që varion në mënyrë progresive nga 2 në 10. Përfundimisht me anë të indekseve të verifikimit kemi përzgjedhur numrin optimal të grupimeve dhe ndarjen optimale në grupime. Në rastin konkret numri optimal i grupimeve ishte pesë dhe ndarja optimale u përftua nga algoritmi i turbullt bazuar në bërthama. Grupimet e përftuara u emërtuan manualisht si "Shumë mirë", " Mirë", "Mesatar", "Mjaftueshëm" dhe "Dobët". Për secilin student vlera e përcaktuar e anëtarësisë së tij në secilin prej këtyre grupimeve (kategorive) është një karakterizues i rëndësishëm i përformancës së tij. Në tabelën vijuese kemi pasqyruar rezultatet për disa studentë më anëtarësinë përkatëse në secilin prej grupimeve.

Numri rendor	Shumë mirë	Mirë	Mesatar	Mjaftueshëm	Dobët
1	0.0106	0.0512	0.8681	0.0610	0.0091
2	0.0326	0.6348	0.2075	0.0784	0.0467
3	0.0012	0.0074	0.0368	0.0521	0.9025
4	0.7645	0.1204	0.0654	0.0349	0.0148
5	0.0085	0.0161	0.0912	0.5718	0.3124

6	0.0215	0.2140	0.6123	0.1352	0.0170
7	0.0124	0.0732	0.7921	0.1139	0.0084
8	0.2735	0.5518	0.1276	0.042	0.0051
9	0.0103	0.1083	0.5914	0.2715	0.0185
10	0.0028	0.0306	0.0816	0.3538	0.5312
11	0.0109	0.0243	0.2154	0.6210	0.1284
12	0.1931	0.6704	0.0822	0.0475	0.0068
13	0.7105	0.2462	0.0434	0.0123	0.0076
14	0.0037	0.0105	0.0320	0.1432	0.8106
15	0.0073	0.03165	0.9135	0.04215	0.0054

Tabelë 12 - Vlerat e anëtarësisë së disa studentëve në grupimet e gjeneruara

Një ndarje e turbullt ofron më shumë informacion për mentorët se sa do të ofronte një ndarje klasike (e fortë/e prerë). Për shembull për studentët e paraqitur përkatësisht me numër rendor 5 dhe numër rendor 11 në rast të një grupimi klasik natyrisht që të dy do të ishin anëtarë të plotë të grupimit "Mjaftueshëm", pa u ofruar ndonjë informacion tjetër shtesë. Ndërkohë në ndarjen e turbullt që kemi paraqitur vërtet të dy studentët kanë anëtarësi më të lartë në gupimin "Mjaftueshëm", por studenti me numër 5 do të këshillohej se është i rrezikuar të përfundojë në katëgorinë "Dobët", ndërsa studenti me numër 11 do të inkurajohej për të arritur grupimin "Mesatar". Në mënyrë të ngjashme studentët me numra 2 dhe 12 që të dy kanë anëtarësinë më të lartë në grupimin "Mirë", por nëse do t'i krahasonim me njëri-tjetrin studenti me numër 12 është më pranë grupimit "Shumë mirë" ndërsa studenti me numër 2 është më pranë grupimit "Mesatar".

4.4.3 Krijimi i profileve karakterizues të performancës specifike

Siç përmendëm edhe në seksionin e kaluar, një nga pjesët kritike të detyrës së një mentori është dhënia e rekomandimeve në lidhje me lëndët me zgjedhje. Në këtë pjesë te dytë të studimit tonë ne synojmë t'i ofrojmë mentorëve informacion ku mund të bazohen për t'i dhënë një përgjigje sa më të specializuar (pra në varësi të studentit) kësaj çështjeje. Përqasja jonë është që rekomandimet për studentët të jepen në bazë të profileve karakterizues të performancës në gruplëndë të caktuara që kanë lidhje me lëndët të cilat do të zgjidhen. Në shembullin konkret ne synojmë t'u japim rekomadime studentëve për të bërë zgjedhjen midis dy lëndëve "Sisteme të shpërndara " dhe "Programimi paralel". Për të analizuar grupimet informuese në lidhje me lëndën "Sisteme të shpërndara" ne përdorim teknikën e shkrirjes (agregimit) mbi të dhënat e tre lëndëve të mëparshme: "Arkitekture kompjuteri", "Sisteme shfrytëzimi" dhe "Rrjete kompjuterike", ndërsa për të analizuar grupimet informuese

në lidhje me lëndën "Programimi paralel" përsëri përdorim teknikën e shkrirjes, por tashme mbi të dhënat e lëndëve "Programim", "Matematikë diskrete" dhe "Analiza e algoritmeve".

Pas etapës së shkrirjes (e konsideruar si një etapë e përpunimit paraprak), vijojmë me aplikimin e algoritmeve të turbullt të grupimit ku përsëri kemi përzgjedhur algoritmin e turbullt të mesoreve, algoritmin Gustafson-Kessel, algoritmin posibilistik të mesoreve dhe algoritmin e turbullt të bazuar në bërthama. Secili algoritëm është ekzekutuar disa herë me vlera të grupimeve që varionin në mënyrë progressive nga 2 në 8. Përfundimisht me anë të indekseve të verifikimit përcaktuam numrin optimal të grupimeve dhe ndarjen optimale në grupime. Në rastin konkret numri optimal i grupimeve ishte katër dhe ndarja optimale u përftua nga algoritmi Gustafson-Kessel. Grupimet e përftuara u emërtuan manualisht si "Shumë mirë", " Kënaqshëm", "Mjaftueshëm" dhe "Dobët". Për secilin student vlera e përcaktuar e anëtarësisë së tij në secilin prej këtyre grupimeve (kategorive) është një karakterizues i rëndësishëm i përformancës së tij në gruplëndën përkatëse. Në tabelën vijuese kemi pasqyruar rezultatet për disa studentë me anëtarësinë përkatëse në secilin prej grupimeve në lidhje me gruplëndën e parë ("Arkitekture kompjuteri", "Sisteme shfrytëzimi" dhe "Rrjete kompjuterike").

Numri rendor	Shumë mirë	Kënaqshëm	Mjaftueshëm	Dobët
1	0.3106	0.5712	0.1053	0.0129
2	0.0609	0.1639	0.6427	0.1325
3	0.8204	0.1275	0.0473	0.0048
4	0.0451	0.9210	0.0211	0.0128
5	0.0247	0.1055	0.7742	0.0956

Tabelë 13 - Vlerat e anëtarësisë për disa studentë në kategoritë e gruplëndës së parë

Në tabelën 14 janë pasqyruar rezultatet e të njejtëve studentë me anëtarësinë përkatëse në secilin prej grupimeve në lidhje me gruplëndën e dytë ("Programim", "Matematikë diskrete" dhe "Analiza e algoritmeve").

Numri rendor	Shumë mirë	Kënaqshëm	Mjaftueshëm	Dobët
1	0.0496	0.1922	0.6357	0.1225
2	0.7193	0.2154	0.0586	0.0067
3	0.8926	0.0685	0.0373	0.0016
4	0.2320	0.6512	0.0854	0.0314
5	0.0865	0.8012	0.0954	0.0169

Tabelë 14 - Vlerat e anëtarësisë për disa studentë në kategoritë e gruplëndës së dytë

Versionet e plota të tabelës 13 dhe 14 u janë dhënë mentorëve për t'i udhëzuar ata në procesin e dhënies së rekomandimeve në lidhje me lëndët me zgjedhje "Sisteme të

shpërndara” dhe “Programimi paralel”. Në rastin konkret për studentin e parë shihet se ka një anëtarësi të lartë në grupimin “Kënaqshëm” dhe grupimin “Shumë mirë” në gruplëndën e parë, si dhe një anëtarësi të lartë në grupimi mjaftueshëm në gruplëndën e dytë. Natyrisht që për këtë student mentori do t’i sugjeronte të zgjedhë lëndën “Sisteme të shpërndara”. Në mënyrë krejt të ngjashme për studentët e dytë dhe të pestë mentori do të sugjeronte zgjedhjen e lëndës “Programimi paralel”, ndërsa për studentët e tretë dhe të katërt nuk kemi një informacion ku të bazohemi ndaj nuk do të jepej sugjerim. Pra sigurisht nuk pretendojmë që të ekzistojë rekomandim për çdo rast, por gjithesi në mjaft raste grupimet e gjeneruara janë mjaft informuese.

5. PËRSHTATJA E ALGORIMIT TË TURBULLT TË MESOREVE PËR NJË PËRQASJE TË MËSIMIT GJYSMË-TË MBIKËQYRUR

Metodologjitë e analizimit të të dhënave (siç e kemi përmendur dhe në fillimin e kapitullit te tretë) i klasifikojmë në dy kategori kryesore: në mësimin e mbikëqyrur dhe mësimin e pambikëqyrur. Teknikat e mësimit të mbikëqyrur veprojnë në kushte ku kemi informacion rreth një pjese të konsiderueshme të të dhënave (p.sh. mund të kemi si informacion të dhëna të kategorizuara, shpërndarje statistikore të të dhënave etj), të cilin e shfrytëzojmë për të realizuar parashikime (klasifikime, kategorizime) për të dhëna të reja. Nga ana tjetër teknikat e mësimit të pambikëqyrur veprojnë në kushte ku nuk ka ndonjë model të specifikuar paraprakisht rreth të dhënave, por synohet të zbulohen struktura të brendshme duke i organizuar të dhënat në grupime bazuar vetëm në ngjashmëritë ndërmjet elementëve individualë. Gjatë kapitullit të tretë dhe të katërt kemi trajtuar si teorikisht dhe praktikisht teknika të analizës (klasike dhe të turbullt) të grupimeve e cila është një nga format më të rëndësishme të mësimit të pambikëqyrur, ndërsa përfshirja në trajtime të detajuara të teknikave të mësimit të mbikëqyrur është përtej qëllimeve të këtij disertacioni.

Nga ana tjetër, në probleme të botës reale shpeshherë mund të jemi në skenare ku ofrohet një sasi e vogël informacioni si për shembull disa elementë të kategorizuar (që janë një pjesë e vogël e tërësisë së elementëve në bashkësinë e të dhënave) ose disa kufizime në formën e çifteve elementësh për të cilët dihet që i përkasin ose nuk i përkasin të njëjtit grupim (kategori). Pra nuk jemi as në një situatë të mungesës së plotë të informacionit paraprak, as në një situatë të prezencës së një informacioni të plotë. Në raste të tilla është e qartë se nuk është e mundur të veprohet bazuar në teknika të mësimit të mbikëqyrur, por gjithashtu veprimi me teknika të mësimit të pambikëqyrur duke e lënë të pashfrytëzuar sasinë e vogël të informacionit të ofruar, nuk do të ishte një zgjedhje optimale. Për të vepruar në mënyrë sa më optimale në rrethana të tilla janë zhvilluar teknikat e mësimit gjysmë-të mbikëqyrur, ideja qëndrore e të cilave është që informacioni (i pakët) që disponohet, të udhëzojë procesin e ndarjes së të dhënave në grupime [49].

Si skenare tipikë ku është i përshtatshëm përdorimi i mësimit gjysmë-të mbikëqyrur mund të përmenden [50]:

- Në diagnozën mjekësore ku do të ishte mjaft e vështirë dhe e kushtueshme të kategorizoheshin të gjithë të dhënat mjekësore (si rrjedhim vetëm një pjesë e të dhënave është e kategorizuar).
- Në përpunimin e imazheve në rastet kur vetëm disa objekte apo rajone të imazhit janë të kategorizuara.
- Në filtrimin e informacionit në rastet kur njohuria rreth elementëve që do të filtrohen është e kufizuar.

- Në motorët e kërkimit (në web) ku kategorizimi i të gjithë dokumenteve apo faqeve do të ishte një detyrë e jashtëzakonshme (pra vetëm një pjesë e dokumenteve/faqeve janë të kategorizuara).
- Në bioinformatikë (në aplikime në lidhje me gjenetikën) ku njohuritë rreth gjeneve përgjegjës për funksionalitete apo sëmundje të caktuara është e kufizuar.

Në dy seksionet vijuese do të trajtojmë teknika të përshtatjes së algoritmit të turbullt të mesoreve për të realizuar një përqasje të mësimit gjysmë-të mbikëqyrur. Në rastin e parë si informacion shtesë kemi disa elementë të kategorizuar, pra elementë të cilëve u njihen kategoritë (grupimet, klasat) paraprakisht. Në rastin e dytë si informacion shtesë kemi disa kufizime në lidhje me çifte elementësh për të cilët dihet nëse janë apo jo në të njëjtin grupim. Për të dyja rastet metodologjia është e njëjtë: modifikimi i funksionit objektiv që synohet të minimizohet. Siç kemi parë gjatë kapitullit të tretë, funksioni objektiv i algoritmit të turbullt të mesoreve shpreh konceptin e shpërhapjes së elementëve brenda grupimeve, pra sa më kompakte të jenë grupimet, aq më e vogël (pra aq më optimale) do të jetë vlera e funksionit objektiv. Ndërsa në rastin e mësimit gjysmë-të mbikëqyrur funksioni objektiv modifikohet duke i shtuar një term të ri i cili shpreh konceptin e "ndotjes" së grupimit. Me "ndotje" të grupimit në rastin e parë nënkuptohet sasia e të dhënave të kategorizuara të cilat nuk janë vendosur në grupimet (kategoritë, klasat) përkatëse, ndërsa në rastin e dytë nënkuptohet sasia e kufizimeve të cilat nuk janë respektuar. Pra funksioni objektiv është një kombinim linear i shpërhapjes së elementëve në grupime dhe i "ndotjes" në grupime.

5.1 Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur i bazuar në të dhëna të kategorizuara

Në këtë seksion do të trajtojmë teknika të modifikimit të funksionit objektiv të algoritmit të turbullt të mesoreve për ta përshtatur të veprojë në mënyrë sa më optimale në prani të një informacioni shtesë që në këtë rast nënkupton njohjen paraprake të grupimeve për disa të dhëna (pra disponohen disa të dhëna të kategorizuara). Ekzistojnë disa teknika modifikimi të funksionit objektiv ku ndër më e njohurat (të cilën e kemi përzgjedhur për trajtimet eksperimentale në seksionin 5.3.1) është forma [51]:

$$J(X,U,C) = \sum_{i=1}^{k}\sum_{j=1}^{n} \mu_{ij}^{\varphi}\, d^2(x_j, c_i) + \alpha \sum_{i=1}^{k}\sum_{j=1}^{n} \left|\mu_{ij} - f_{ij} b_j\right|^{\varphi} d^2(x_j, c_i)$$

Në shënimin e mësipërm n është numri i elementëve në bashkësinë e të dhënave, x_j është elementi i j-të në bashkësinë e të dhënave, k është numri i grupimeve, U është matrica particion ku çdo vlerë e saj μ_{ij} përfaqëson anëtarësinë në grupimin e i-të të

elementit të j-të, c_i është qendra e grupimit të i-të, $d^2(x_j, c_i)$ është katrori i distancës midis x_j dhe c_i, φ është koeficienti i turbullsisë, f_{ij} është anëtarësia në grupimin e i-të të elementit të j-të të kategorizuar, b_j është një vlerë binare (pra 0 ose 1) ku vlera 1 tregon se elementi x_j është i kategorizuar, ndërsa 0 tregon se elementi x_j nuk është i kategorizuar. Së fundmi α është një parametër jonegativ që kontrollon ndikimin e të dhënave të kategorizuara në procesin e grupimit (sa më e madhe vlera e α, aq më i madh është ndikimi). Në rastin e veçantë kur $\alpha = 0$ funksioni objektiv është i njëjtë me algoritmin e turbullt të mesoreve. Gjithashtu një aspekt për t'u vënë në dukje i kësaj përqasjeje është se numri i grupimeve supozohet i parapërcaktuar, duke shmangur kështu një nga çështjet kritike të algoritmit të turbullt të mesoreve.
Për optimizimin e funksionit objektiv, në mënyrë të ngjashme me algoritmin e turbullt të mesoreve, përdoret një skemë iteracioni. Natyrisht që si rrjedhim i modifikimit që kryem mbi funksionin objektiv ndryshon mënyra si do të përditësohet matrica particion (hapi i katërt në pseudokodin e mëposhtëm) dhe përllogariten vlerat e qendrave të reja (hapi i pestë në pseudokodin e mëposhtëm). Në vijim paraqesim pseudokodin që përshkruan hapat kryesorë të algoritmit të turbullt të mesoreve gjysmë-të mbikëqyrur bazuar në të dhëna të kategorizuara [51]:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. U jepet vlera fillestare (zakonisht 0 ose rastësisht) të gjithë elementëve të matricës U_0.
3. Llogaritet distanca ndërmjet secilit element dhe secilës qendër të grupimeve sipas funksionit distancë (preferohet distanca Mahalanobis).
4. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, sipas:

$$\mu_{ij} = \frac{1}{1+\alpha}\left(\frac{1+\alpha(1-b_j\sum_{t=1}^{k} f_{tj})}{\sum_{t=1}^{k}\left(\frac{d_{ij}}{d_{tj}}\right)^2} + \alpha f_{ij} b_j\right)$$

5. Llogariten qendrat e reja të grupimeve:

$$c_i = \frac{\sum_{j=1}^{n} {\mu_{ij}}^2 x_j}{\sum_{j=1}^{n} {\mu_{ij}}^2} \text{ për çdo } i \in \{1,2,\dots,k\}$$

6. $nr = nr + 1$ (rritet me një numri i iteracioneve)
7. Nqs $||U_{nr} - U_{nr-1}|| > \varepsilon$ atëherë vazhdo në hapin 2.
8. FUND

Figurë 28 - Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur bazuar në të dhëna të kategorizuara

Përveç formës së lartpërmendur ekzistojnë dhe forma të tjera si mund të modifikohet algoritmi i turbullt i mesoreve për një përqasje të mësimit të mbikëqyrur. Në vijim do të përshkruajmë në mënyrë të përmbledhur tri metoda të tjera, ku për secilën do të paraqesim si modifikohet funksioni objektiv, si përditësohet matrica particion dhe si përllogariten qendrat e reja të grupimeve. Për secilën nga këto metoda pseudokodi përshkrues i algoritmit mund të formulohet si në figurën 28 me ndryshimet përkatëse në mënyrën e përditesimit të matricës particion (hapi i katërt) dhe në mënyrën e përllogaritjes së qendrave të reja të grupimeve (hapi i pestë).

- Mënyra e parë [52]:

  Funksioni objektiv: $J(X,U,C) = \sum_{i=1}^{k}\sum_{j=L+1}^{n}\mu_{ij}^{\varphi}\, d^2(x_j,c_i) +$

  $$(1-\alpha)\sum_{i=1}^{k}\sum_{j=1}^{L}\mu_{ij}^{\varphi}\, d^2(x_j,c_i) + \alpha\sum_{i=1}^{k}\sum_{j=1}^{n}\left|\mu_{ij} - f_{ij}b_j\right|^{\varphi} d^2(x_j,c_i)$$

  Përditësimi i matricës particion:

  $$\mu_{ij} = \begin{cases} \dfrac{1}{\sum_{t=1}^{k}\left(\dfrac{d_{ij}}{d_{tj}}\right)^2} & \text{nqs } x_j \text{ nuk është e kategorizuar} \\ \alpha f_{ij} + \dfrac{1}{\sum_{t=1}^{k}\left(\dfrac{d_{ij}}{d_{tj}}\right)^2} & \text{nqs } x_j \text{ është e kategorizuar} \end{cases}$$

  Përllogaritja e qendrave të reja:

  $$c_i = \frac{\sum_{j=1}^{n}\mu_{ij}{}^2 x_j}{\sum_{j=1}^{n}\mu_{ij}{}^2} \text{ për çdo } i \in \{1,2,\dots,k\}$$

- Mënyra e dytë [52]:

  Funksioni objektiv:

  $$J(X,U,C) = 2\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}\left(1 - \mathrm{K}(x_i,c_j)\right)$$

  Përditësimi i matricës particion:

  $$\mu_{ij}^{u} = \frac{(1/(1-K(x_j^u,c_i)))^{1/(\varphi-1)}}{\sum_{t=1}^{k}(1/(1-K(x_j^u,c_t)))^{1/(\varphi-1)}}$$

  Përllogaritja e qendrave të reja:

  $$c_i = \frac{\sum_{j=1}^{n_1}(\mu_{ij}^{l})^{\varphi} x_j\, K(x_j^l,c_i)x_j^l + \sum_{j=1}^{n_2}(\mu_{ij}^{u})^{\varphi} x_j\, K(x_j^u,c_i)x_j^u}{\sum_{j=1}^{n_1}(\mu_{ij}^{l})^{\varphi} x_j\, K(x_j^l,c_i) + \sum_{j=1}^{n_2}(\mu_{ij}^{u})^{\varphi} x_j\, K(x_j^u,c_i)}$$

  për çdo $i \in \{1,2,\dots,k\}$, ku n_1 është numri i të dhënave të kategorizuara, n_2 është numri i të dhënave të pakategorizuara, me mbishënimin l nënkuptojmë një të dhënë të kategorizuar ("labeled"), ndërsa me mbishënimin u nënkuptojmë një të dhënë të pakategorizuar ("unlabelled").

- Mënyra e tretë [52]:
  Funksioni objektiv:

$$J(X,U,C)=\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{\mathrm{ij}}\,d^2\left(x_j,c_i\right)+\lambda^{-1}\sum_{i=1}^{k}\sum_{j=1}^{L}\left(\mu_{\mathrm{ij}}-\bar{\mu}_{\mathrm{ij}}\right)\log\left|\mu_{\mathrm{ij}}-\bar{\mu}_{\mathrm{ij}}\right|$$

  Përditësimi i matricës particion:

$$\mu_{\mathrm{ij}}=\bar{\mu}_{\mathrm{ij}}+\frac{\mathrm{e}^{-\lambda d\left(x_j,c_i\right)}}{\sum_{t=1}^{k}\mathrm{e}^{-\lambda d\left(x_j,c_t\right)}}$$

  Përllogaritja e qendrave të reja:

$$c_i=\frac{\sum_{j=1}^{n}\mu_{ij}\,x_j}{\sum_{j=1}^{n}\mu_{ij}}$$

5.2 Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur i bazuar në kufizime

Në këtë seksion do të trajtojmë teknika të modifikimit të funksionit objektiv të algoritmit të turbullt të mesoreve për ta përshtatur të veprojë në mënyrë sa më optimale në prani të një informacioni shtesë që në këtë rast nënkupton disa kufizime në lidhje me çifte të dhënash për të cilat dihet paraprakisht nësë i përkasin të njëjtit grupim apo jo (por nuk dihet se cilit grupim i përkasin këto të dhëna). Informacionin shtesë që disponohet, e shprehim në mënyrë të strukturuar si dy bashkësi çiftesh të renditura R_1 dhe R_2. Bashkësia R_1 përmban çifte të renditura elementësh të bashkësisë së të dhënave, ku $\left(x_i,x_j\right)\in R_1$ do të thotë se elementët x_i dhe x_j duhet të ndodhen në të njëjtin grupim në ndarjen përfundimtare. Nga ana tjetër bashkësia R_2 përmban çifte të renditura elementësh të bashkësisë së të dhënave, ku $\left(x_i,x_j\right)\in R_2$ do të thotë se elementët x_i dhe x_j nuk duhet të ndodhen në të njëjtin grupim në ndarjen përfundimtare. Metodologjia e përdorur edhe në këtë rast është modifikimi i funksionit objektiv për marrjen në konsideratë të informacionit të disponuar. Ekzistojnë disa teknika modifikimi ku ndër më e njohurat (të cilën e kemi përzgjedhur për trajtimet eksperimentale në seksionin 5.3.2) është forma [53]:

$$J(X,U,C)=\sum_{i=1}^{k}\sum_{j=1}^{n}\mu_{ij}^{\varphi}\,d^2\left(x_j,c_i\right)+\alpha\left(\sum_{(x_i,x_j)\in R_1}\sum_{s=1}^{k}\sum_{t=1,l\neq k}^{k}\mu_{\mathrm{si}}\mu_{\mathrm{tj}}+\sum_{(x_i,x_j)\in R_2}\sum_{t=1}^{k}\mu_{\mathrm{ti}}\mu_{\mathrm{tj}}\right)$$

Në shënimin e mësipërm kemi përdorur saktësisht të njëjtat konvencione si më parë (në seksionin 5.1) për n, x_j, μ_{ij}, $d^2\left(x_j,c_i\right)$ dhe φ. Gjithashtu roli i parametrit α është

shumë i ngjashëm, pra të kontrollojë ndikimin e kufizimeve në procesin e grupimit (sa më e madhe vlera e α, aq më i madh është ndikimi). Në rastin e veçantë kur $\alpha = 0$ funksioni objektiv është i njëjtë me algoritmin e turbullt të mesoreve. Siç shihet, termi shtesë në funksionin objekiv përbëhet prej dy shumash ku e para ka lidhje me me elementët të cilët duhet të jenë në të njëjtin grupim, ndërsa e dyta ka të bëjë me elementët të cilët nuk duhet të jenë në të njëjtin grupim.

Për optimizimin e funksionit objektiv, në mënyrë të ngjashme me algoritmin e turbullt të mesoreve, përdoret një skemë iteracioni. Natyrisht që si rrjedhim i modifikimit që kryem mbi funksionin objektiv ndryshon mënyra si do të përditësohet matrica particion (hapi i katërt në pseudokodin e mëposhtëm) dhe përllogariten vlerat e qendrave të reja (hapi i pestë në pseudokodin e mëposhtëm). Në vijim paraqesim pseudokodin që përshkruajn hapat kryesorë të algoritmit të turbullt të mesoreve gjysmë-të mbikëqyrur bazuar në kufizime:

1. Shënojmë $nr = 1$ (numëruesi i iteracioneve)
2. U jepet vlera fillestare (të shpërndara uniformisht) të gjithë elementëve të matricës U_0.
3. Llogaritet distanca ndërmjet secilit element dhe secilës qendër të grupimeve sipas funksionit distance.
4. Përditësohet matrica particion $U_{nr} = [\mu_{ij}]$, me vlera

$$\mu_{ij} = (\mu_{ij})_{ATM} + (\mu_{ij})_{kufizime}$$

Ku $(\mu_{ij})_{ATM} = \frac{1}{\sum_{t=1}^{k}\left(\frac{d_{ij}}{d_{tj}}\right)^2}$ dhe $(\mu_{ij})_{kufizime} = \frac{\alpha\left(\bar{c}_{x_i} - c_{x_j}\right)}{d^2(x_j, c_i)}$

5. Llogariten qendrat e reja të grupimeve:

$c_i = \frac{\sum_{j=1}^{n} \mu_{ij}{}^2 x_j}{\sum_{j=1}^{n} \mu_{ij}{}^2}$ për çdo $i \in \{1,2,\dots,k\}$

6. $nr = nr + 1$ (rritet me një numri i iteracioneve)
7. Nqs $||U_{nr} - U_{nr-1}|| > \varepsilon$ atëherë vazhdo në hapin 2.
8. FUND

Figurë 29 - Algoritmi i turbullt i mesoreve gjysmë-i mbikëqyrur bazuar në kufizime

5.3 Studime eksperimentale mbi algoritmet e grupimit të turbullt gjysmë-të mbikëqyrur

Në këtë seksion do të zhvillojmë disa studime eksperimentale mbi teknikat e mësimit gjysmë-të mbikëqyrur që trajtuam në dy seksionet e mëparshme. Për të realizuar procedurat eksperimentale përsëri do të shfrytëzojmë disa bashkësi të dhënash standarde të ofruara publikisht në depozitën e të dhënave të Universitetit të Kalifornisë në Irvine, si dhe një bashkësi të dhënash sintetike. Bashkësitë e të dhënave standarde që kemi përzgjedhur janë Glass, Dermatology, Breast Cancer Wisconsin dhe Yeast, ndërsa nga bashkësitë sintetike kemi përzgjedhur Synth2. Për këto bashkësi struktura e brendshme (grupimet e brendshme) është e njohur paraprakisht. Kur bashkësitë në fjalë i përdorëm në procedurat eksperimentale në kapitullin e katërt, ne shfrytëzuam të gatshëm vetëm numrin e grupimeve. Grupimet e njohura të këtyre bashkësive nuk u shfrytëzuan aspak gjatë procesit të analizës së grupimeve, por u përdorën vetëm në fund për të vlerësuar shkallën e saktësisë së algoritmeve përkatës. Gjatë procedurave eksperimentale të këtij seksioni ne nuk do të shfrytëzojmë të gatshëm vetëm numrin e grupimeve, por edhe infomacione shtesë. Në eksperimentet e seksionit 5.3.1 përveç numrit të grupimeve, do të përzgjedhim për të shfrytëzuar të gatshëm kategoritë (grupimet) e disa prej elementëve. Përzgjedhja e këtyre elementëve është realizuar në mënyrë të rastësishme dhe në mënyrë progresive janë marrë 3%, 8%, 12% , 15% dhe 20% e elementëve. Në eksperimentet e seksionit 5.3.2 përveç numrit të grupimeve, do të përzgjedhim për të shfrytëzuar të gatshëm disa kufizime në lidhje me të dhëna të cilat duhet të jenë ose jo në të njëjtin grupim. Përzgjedhja e këtyre kufizimeve është realizuar në mënyrë të rastësishme dhe numrin e kufizimeve e variojmë në mënyrë progresive duke marrë përkatësisht 10, 20, 30, 40 dhe 50 kufizime.

5.3.1 Eksperimente mbi algoritmin e turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në të dhëna të kategorizuara

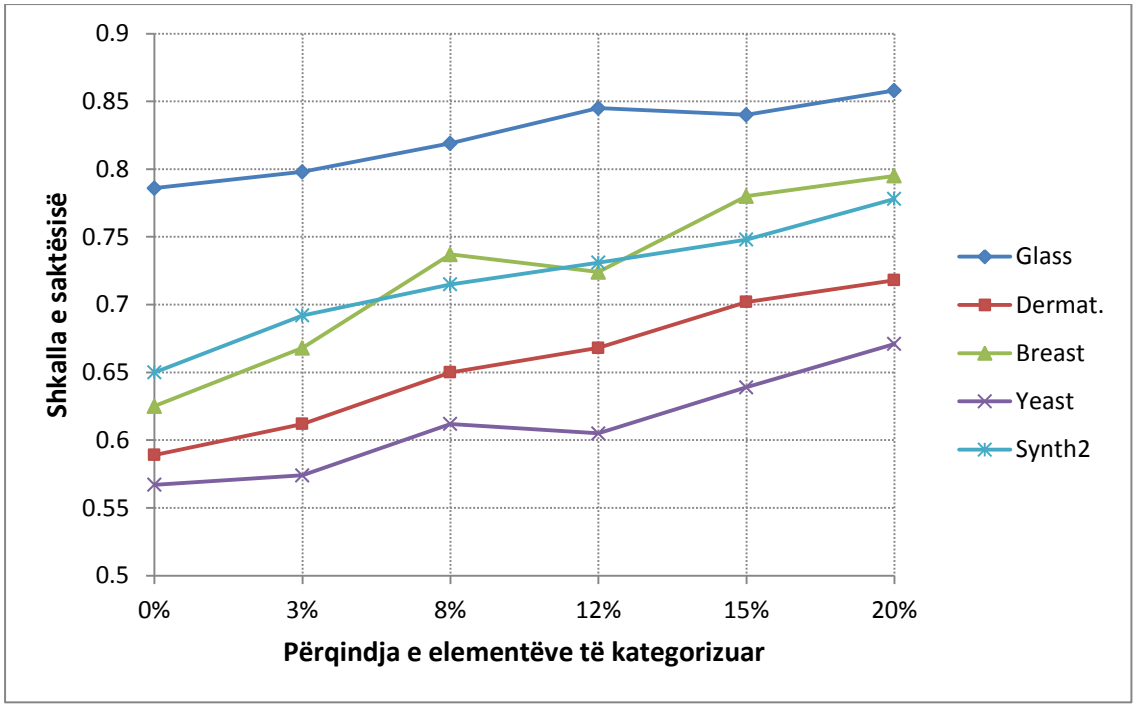
Në eksperimentet e këtij seksioni kemi vënë në përdorim bashkësitë e të dhënave Glass, Dermatology, Breast Cancer Wisconsin, Yeast dhe Synth2. Për secilën prej tyre shfrytëzojmë të gatshëm numrin e grupimeve, si dhe një pjesë të dhënash të kategorizuara, të cilat janë përzgjedhur në mënyrë të rastësishme. Sasia e të dhënave është zgjedhur në raport me numrin total të elementëve në bashkësi. Eksperimentet janë kryer me vlera të këtij raporti përkatësisht 3%, 8%, 12% , 15% dhe 20%. Për të dhënë një përshkrim krahasues sa më të plotë kemi zhvilluar eksperimente edhe me raportin 0%, por në këtë rast është e qartë se nuk kemi të bëjmë me mësim gjysmë-të mbikëqyrur, por thjesht me algoritmin e turbullt të mesoreve (i cili është pjesë e

mësimit të pambikëqyrur). Në fund të eksperimenteve, rezultatet e gjeneruara nga eksperimenti përkatës i krahasojmë me grupimet e njohura të secilës bashkësi për të përllogaritur shkallën e saktësisë. Rezultatet e eksperimenteve i përmbledhim në tabelën e mëposhtme:

Bashkësia e të dhënave	A.T.M	A.T.M gjysmë i mbikëqyrur bazuar në të dhëna të kategorizuara				
	0%	3%	8%	12%	15%	20%
Glass	0.786	0.798	0.819	0.845	0.840	0.858
Dermatology	0.589	0.612	0.650	0.668	0.702	0.718
Breast Cancer Wisconsin	0.625	0.668	0.737	0.724	0.78	0.795
Yeast	0.567	0.574	0.612	0.605	0.639	0.671
Synth2	0.65	0.692	0.715	0.731	0.748	0.778

Tabelë 15- Shkalla e saktësisë së algoritmit në varësi të përqindjes së elementëve të kategorizuar

Rezultatet e këtij eksperimenti gjithashtu mund t'i paraqesim vizualisht me anë të grafikut si në vijim. Për secilën nga pesë bashkësitë paraqesim si ndryshon shkalla e saktësisë së algoritmit të turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në të dhëna të kategorizuara, në varësi të raportit (përqindjes) së elementëve të kategorizuar.



Figurë 30- Shkalla e saktësisë së algoritmit në varësi të përqindjes së elementëve të kategorizuar

Nga tabela dhe grafiku i paraqitur vihet re qartë se përqasja e mësimit gjysmë-të mbikëqyrur për algoritmin e turbullt të mesoreve, e përmirëson ndjeshëm shkallën e saktësisë së tij. Edhe për sasi të vogla të përqindjes së elementëve të kategorizuar, ka përmirësim në shkallën e saktësisë. Një tjetër vëzhgim i rëndësishëm është se në përgjithësi me rritjen e përqindjes së elementëvë të kategorizuar, rritet dhe shkalla e saktësisë së algoritmit, por gjithashtu në raste të rralla rritja e përqindjes së elementëve të kategorizuar mund të mos e përmirësojë shkallën e saktësisë (madje mund ta përkeqesojë lehtas atë).

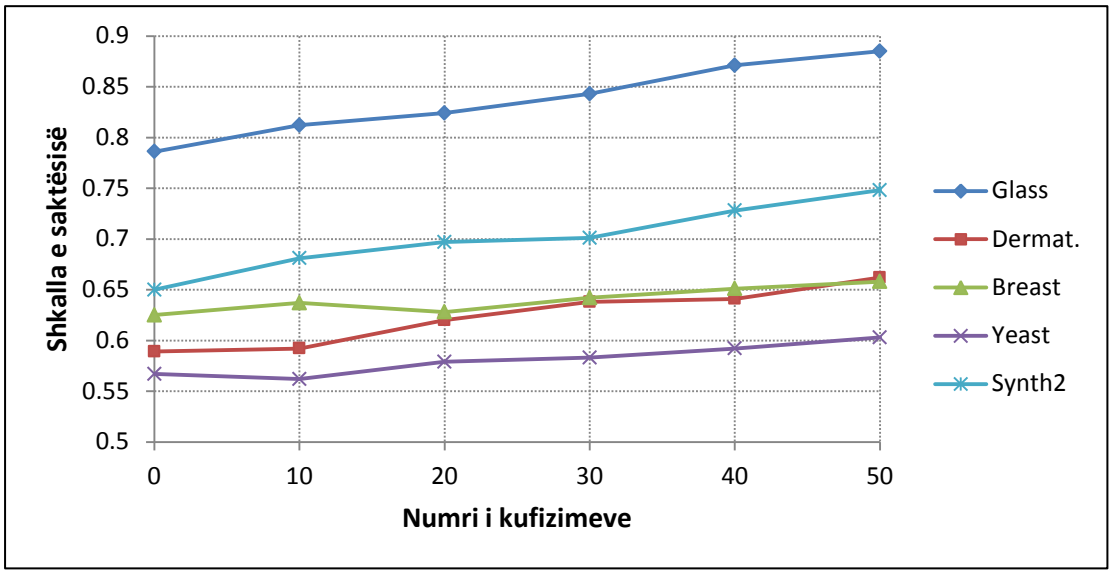
5.3.2 Eksperimente mbi algoritmin e turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në kufizime

Edhe në eksperimentet e këtij seksioni kemi vënë në përdorim bashkësitë e të dhënave Glass, Dermatology, Breast Cancer Wisconsin, Yeast dhe Synth2. Për secilën prej tyre shfrytëzojmë të gatshëm numrin e grupimeve, si dhe disa kufizime në lidhje me çifte elementësh të bashkësisë së të dhënave të cilët duhet të jenë ose jo në të njëjtin grupim. Për secilën nga bashkësitë kemi kryer eksperimente me kufizime të përzgjedhur në mënyrë të rastësishme, ku numri i tyre varion në 10, 20, 30, 40 dhe 50. Për të dhënë një përshkrim krahasues sa më të plotë kemi zhvilluar eksperimente edhe me asnjë (zero) kufizim por në këtë rast është e qartë se nuk kemi të bëjmë me mësim gjysmë-të mbikëqyrur, por thjesht me algoritmin e turbullt të mesoreve (i cili është pjesë e mësimit të pambikëqyrur). Në fund të eksperimenteve, rezultatet e gjeneruara nga eksperimenti përkatës i krahasojmë me grupimet e njohura të secilës bashkësi për të përllogaritur shkallën e saktësisë. Rezultatet e eksperimenteve i përmbledhim në tabelën e mëposhtme:

Bashkësia e të dhënave	**A.T.M**	**A.T.M gjysmë i mbikëqyrur bazuar në kufizime**				
	0	**10**	**20**	**30**	**40**	**50**
Glass	0.786	0.812	0.824	0.843	0.871	0.885
Dermatology	0.589	0.592	0.620	0.638	0.641	0.662
Breast Cancer Wisconsin	0.625	0.637	0.628	0.642	0.651	0.658
Yeast	0.567	0.562	0.579	0.583	0.592	0.603
Synth2	0.65	0.681	0.697	0.701	0.728	0.748

Tabelë 16 - Shkalla e saktësisë së algoritmit në varësi të numrit të kufizimeve

Rezultatet e këtij eksperimenti gjithashtu mund t'i paraqesim vizualisht me anë të grafikut si në vijim. Për secilën nga pesë bashkësitë paraqesim si ndryshon shkalla e saktësisë së algoritmit të turbullt të mesoreve gjysmë-të mbikëqyrur të bazuar në të kufizime, në varësi të numrit të kufizimeve të përzgjedhur.



Figurë 31 - Shkalla e saktësisë së algoritmit në varësi të numrit të kufizimeve

Nga tabela dhe grafiku i paraqitur vihet re qartë se përqasja e mësimit gjysmë-të mbikëqyrur për algoritmin e turbullt të mesoreve, e përmirëson ndjeshëm shkallën e saktësisë së tij. Në përgjithësi me rritjen e numrit të kufizimeve, rritet dhe shkalla e saktësisë së algoritmit, por gjithashtu në raste të rralla rritja numrit të kufizimeve mund të mos e përmirësojë shkallën e saktësisë (madje mund ta përkeqësojë lehtas atë).

Vihet re se në bashkesinë Glass dhe Synth2 kemi një rritje më të shpejtë të shkallës së saktësisë me rritjen e numrit të kufizimeve. Natyra e këtij eksperimenti ishte e tillë që numrin e kufizimeve e mbajtëm të fiksuar për të gjithë bashkësitë, ndaj efekti ndjehet më shumë për bashkësitë me numër më të vogël të të dhënave (Glass dhe Synth).

6. METODAT ANSAMBËL NË ANALIZËN E TURBULLT TË GRUPIMEVE

Pavarësisht fushës së gjerë të aplikimeve të suksesshme të algoritmeve të grupimit të turbullt, ekzistojnë disavantazhe të rëndësishme për t'u marrë parasysh si ndjeshmëria ndaj përzgjedhjes së vlerave fillestare për qendrat e grupimeve apo ndjeshmëria ndaj prezencës së vlerave të largëta dhe vlerave me zhurmë në bashkësinë e të dhënave. Këto çështje janë kritike për algoritmet e turbullt të grupimeve sepse mund të ndikojnë në uljen e stabilitetit të tyre dhe mund të cënojnë shkallën e saktësisë në grupimet e gjeneruara. Metodat ansambël janë një fushë relativisht e re studimi (fillesat e së cilave i përkasin fundit të viteve '90) dhe synojnë të trajtojnë disavantazhet e lartpërmendura të algoritmeve të grupimit. Studimi i metodave ansambël në problemin e analizës së grupimeve (mësimi i pambikëqyrur) është frymëzuar nga suksesi i metodave ansambël në problemin e klasifikimit (mësimi i mbikëqyrur).

Ideja qëndrore e metodave ansambël është ekzekutimi i disa (në vend të një) algoritmeve grupues mbi bashkësinë e të dhënave duke gjeneruar disa ndarje të të dhënave në grupime dhe më pas ndarjet në fjalë shkrihen (kombinohen me njëra-tjetrën) në një ndarje të vetme. Qëllimi i kësaj metodologjie është përmirësimi i saktësisë dhe stabilitetit në trajtimin e problemit të analizës së turbullt të grupimeve [55]. Vemë në dukje se edhe më parë gjatë procedurave eksperimentale të kapitullit të katërt ne kemi vënë në zbatim idenë e ekzekutimit të disa algoritmeve grupues, por më pas ne thjesht zgjidhnim ndër rezultatet e përftuara ndarjen më cilësore sipas disa indekseve verifikues (pra ndarjet në grupime të gjeneruara nga algoritmet grupues nuk modifikoheshin më tej, nuk kombinoheshin me njëra-tjetrën për të formuar ndarjen përfundimtare, vetëm midis tyre bëhëj përzgjedhja).

Metodat ansambël kanë gjetur zbatime të rëndësishme në një sërë skenaresh të botës reale, ku ndër më kryesoret përmendim [56]:

- *Sistemet e shpërndara*: Të dhënat e një aplikacioni mund të jenë të shpërndara në një sërë pajisjesh të lidhura në një rrjet të madh. Komunikimi i të gjithë këtyre të dhënave nëpërmjet rrjetit për të realizuar analizën e grupimeve nuk do të ishte zgjidhje e përshtatshme jo vetëm prej shkallës së lartë të kompleksitetit njehsues, por edhe për probleme privatësie apo sigurie. Nga ana tjetër në përqajen e metodave ansambël, në rrjet nuk komunikohet bashkësia e të dhënave, por vetëm rezultatet e grupimeve duke shmangur kështu problemet e përmendura më sipër.
- *Ripërdorimi i informacionit*: Gjatë rigrupimit të të dhënave mund të shfrytëzohen grupime të njohura më parë (të trashëguara) në lidhje më këto të dhëna. Megjithëse për këto grupime mund të mos dihet mënyra si ato janë gjeneruar (madje mund të jenë gjeneruar manualisht nga ekspertë të fushës

përkatësi), metodat ansambël mundësojnë shfrytëzimin e tyre në krijimin e grupimeve të reja të bashkësisë së të dhënave.

- *Grupimet nga disa këndvështrime*: Në një bashkësi të dhënash mund të realizohet analiza e grupimeve disa herë duke vënë në punë atribute apo kritere të ndryshme. Për shembull në aplikime të fushës së marketingut klientët mund të grupohen sipas nevojave, profileve demografike, preferencave të markave etj. Metodat ansambël mund të shfrytëzohen për të gjeneruar një ndarje të vetme bazuar në ndarjet e mëparshme.
- *Përmirësimi i cilësisë së rezultateve*: Të gjithë algoritmet e grupimit kanë pikat e tyre të dobëta në lidhje me aftësinë e eksplorimit të bashkësive të të dhënave të natyrave të ndyshme. Kombinimi i rezultateve të algoritmeve të ndryshëm me anë të metoave ansambël ofron përmirësime të dukshme përkundrejt zgjidhjeve individuale, madje studimet kanë treguar se sa më i lartë të jetë diversiteti ndërmjet ndarjeve të ndryshme, aq e lartë është shkalla e saktësisë e mesatares së tyre.

Në pjesën vijuese të këtij kapitulli, fillimisht do të trajtojmë parimet e përgjithshme të projektimit të metodave ansambël (theksojmë se ezkiston një larmi metodologjish për ndërtimin e metodave ansambël), duke vënë në dukje mënyra të realizimit të dy etapave thelbësore: gjenerimi i grupimeve të ndryshëm dhe shkrirja e tyre në grupim përfundimtar. Më tej do të propozojmë dy modele metodash ansambël: një model homogjen dhe një model heterogjen. Në përfundim do të zhvillojmë disa procedura eksperimentale ku të studiojmë dhe krahasojmë shkallën e sakësisë së modeleve ansambël të propozuar, të cilët do t'i përdorim për të shmangur disavantazhin e ndjeshmërisë ndaj vlerave fillestare të qendrave të grupimeve, që karakterizon algoritmet e turbullt të grupimit.

6.1 Parimet e përgjithshme të projektimit të metodave ansambël

Projektimi i një metode ansambël në vija të trasha është i përbërë nga dy etapa: përzgjedhja e algoritmeve grupues individualë (të cilët do të përdoren për të gjeneruar një seri ndarjesh të ndryshme) dhe mënyra si rezultatet e gjeneruara do të shkrihen (kombinohen me njëra-tjetrën) për të formuar ndarjen përfundimtare. Në disa raste këto dy etapa janë të ndërthurura në një të vetme, pra sa herë që shtohet një ndarje e re e gjeneruar nga ekzekutimi i një algoritmi grupues, ndarja totale përditësohet duke marrë në konsideratë edhe ndarjen e re [57]. Gjithsesi në studimin tonë këto dy etapa do t'i trajtojmë të ndara qartë nga njëra-tjetra duke qenë se kështu ofrohet larmishmëri në mënyrat si mund të projektohet një metodë ansambël (pra çdo mënyrë e gjenerimint të ndarjeve mund të kombinohet me çdo mënyrë të shkrirjes së ndarjeve

në ndarjen përfundimtare). Për sa i përket etapës së parë, disa nga variantet kryesore që mund të përzgjedhim janë [58]:

- Ekzekutimi disa herë i të njëjtit algoritëm të analizës së grupimeve me vlera të ndyshme fillestare të qendrave të grupimeve.
- Ekzekutimi disa herë i të njëjtit algoritëm të analizës së grupimeve me vlera të ndyshme të një ose disa parametrave karakterizues të algoritmit si për shembull koeficienti i turbullsisë, funksioni distancë etj. Kjo procedurë dhe procedura e mëparshme njihen si grupimi homogjen (pra algoritmi i ekzekutuar është i njëjtë, ndërkohë që ndyshojnë vlerat fillestare të qendrave të grupimeve apo parametrat).
- Ekzekutimi disa herë i të njëjtit algoritëm të analizës së grupimeve duke u bazuar në bashkësi të ndyshme të atributeve të elementëve të bashkësisë së të dhënave. Kjo procedurë njihet si grupimi i shpërndarë sipas atributeve.
- Ekzekutimi algoritmeve të ndryshëm të analizës së grupimeve. Kjo procedurë njihet si grupimi heterogjen ose grupimi hibrid.
- Ekzekutimi i algoritmeve të grupimit mbi disa nënbashkësi të ndryshme të bashkësisë së të dhënave. Kjo procedurë njihet si grupimi i shpërndarë.

Përveç mënyrave themelore të lartpërmendura, ekzistojnë dhe mënyra të tjera si kombinim i dy apo më shumë prej varianteve, për shembull mund të ekzekutohen disa herë algoritme të ndryshëm dhe njëkohësisht për secilin prej tyre variojmë një ose disa prej parametrave karakterizues si koeficienti i turbullsisë, funksioni distancë etj. Gjithashtu në modelin homogjen që do të propozojmë në vijim, kombinojmë njëkohësisht variantin e parë dhe variantin e dytë për etapën e përzgjedhjes së algoritmit grupues.

Në përfundim të etapës së parë si rezultat kemi një seri ndarjesh të ndryshme, të cilat po i shënojmë $U_1, U_2, \dots, U_r$ dhe tashmë gjatë etapës së dytë të metodave ansambël synimi ynë do të jetë kombinimi i këtyre ndarjeve në një ndarje të vetme përfundimtare $U*$. Për të realizuar këtë proces ekzistojnë një sërë përqasjesh ku ndër më kryesoret janë [58], [59]:

- Përqasja e ri-kategorizimit (e njohur si përqasja e drejtpërdrejtë) përpiqet të zbulojë marrëdhënie midis grupimeve të gjeneruara në ndarje të ndryshme në mënyrë që t'i kombinojë ato në ndarjen përfundimtare. Theksojmë se në kategorizim duhen marrë parasysh dhe rastet e përkëmbimit të grupimeve.
- Përqasjet e bazuara në teorinë e grafeve përdorin koncepte të kësaj teorie për të përcaktuar ndarjen përfundimtare në grupime. Ndër këto përqasje si më të njohura përmendim algoritmin e ndarjes të bazuar në ngjashmërinë e grupimeve (në teori njihet dhe me shkurtimin CSPA), algoritmi i ndarjes i hipergrafeve (njihet me shkurtimin HGPA), algoritmi i meta-grupimeve

(njihet me shkurtimin MCLA) dhe algoritmi i formulimit të grafit dypjesësh hibrid (njhet me shkurtimin HBGF).

- Përqasja e bazuar në atribute (e cila njihet dhe si modeli i përzier) e trajton rezultatin e secilit algoritëm të grupimit si një atribut kategorik. Bashkësia e ndarjeve konceptehet si një bashkësi të dhënash me r elementë mbi të cilën zbatohen një algoritëm grupimi tjetër për të gjeneruar ndarjen përfundimtare.
- Përqasja e shoqërimit e trajton problemin e përcaktimit të korrecpondencës midis grupimeve duke ndërtuar një matricë shoqërimi për secilën ndarje. Kjo matricë përmban të dhënat e përkatësisë në grupim për të gjithë çiftet e elementëve të bashkësisë së të dhënave. Në rastin e grupimeve klasike vlerat në matricë janë ose 1 (nëse të dy elementët i përkasin të njëjtit grupim) ose 0 (nëse elementët i përkasin grupimeve të ndryshme). Nga ana tjetër në rastin e grupimeve të turbullta vlerat e matricës së shoqërimit përcaktohen me anë të t-normave. Më tej këto matrica kombinohen me njëra tjetrën në matricën e përkatësisë (e cila zakonisht përllogaritet si mesatarja aritmetike e të gjithë matricave të shoqërimit).

Ndër këto përqasje, në modelet ansambël që do të propozojmë në seksionet vijuese ne kemi përzgjedhur përqasjen e shoqërimit. Në lidhje me këtë përqasje mund të specifikojme më tej se për sa i përket mënyrës si matricat kombinohen me njëra-tjetrën për të formuar ndarjen përfundimtare në grupime, ekzistojnë dy mënyra kryesore: përdorimi i prerjeve alfa ose përdorimi i një algoritmi grupues i cili i shfrytëzon vlerat e matricave si vlera të funksioneve distancë.

6.2 Një model homogjen i një metode ansambël të grupimit të turbullt

Në këtë seksion do të projektojmë një metodë ansambël homogjene të grupimit të turbullt të cilën do ta implementojmë (në seksionin 6.4) si dhe do të studiojmë eksperimentalisht shkallën e saktësisë së saj. Një metodë ansambël quhet homogjene kur ndarjet gjenerohen nga i njëjti algoritëm, por duke ndryshuar vlerat fillestare të qendrave të grupimeve ose duke ndryshuar një ose disa nga parametrat karkterizues të algoritmit si për shembull koeficienti i turbullsisë, funksioni distancë, shkalla e tolerancës etj.

Algoritmi që do të shfrytëzojmë në këtë metodë për të realizuar etapën e parë të saj (gjenerimi i ndarjeve) është algoritmi i turbullt i mesoreve. Algoritmin në fjalë do ta ekzekutojmë për disa vlera të ndryshme të qendrave fillestare (të gjitha të përzgjedhura në mënyrë të rastësishme). Për secilën bashkësi vlerash të qendrave fillestare algoritmin e ekzekutojmë katër herë duke variuar koeficientin e turbullsisë përkatësisht me vlerat 1.25, 1.5, 2 dhe 2.5.

Më tej në fazën e shkrirjes do të vemë në përdorim përqasjen e shoqërimit. Për secilën ndarje vlerat e matricës së shoqërimit përllogariten me anë të një t-norme, ku ne kemi përzgjedhur t-normën prodhim $T(x, y) = x \cdot y$. Të gjithë matricat e shoqërimit shkrihen më pas në matricën e përkatësisë (e cila përllogaritet si mesatarja aritmetike e të gjithë matricave të shoqërimit). Përfundimisht mbi rreshtat (ose shtyllat) e matricës së përkatësisë zbatojmë përsëri algoritmin e turbullt të mesoreve për të gjeneruar ndarjet përfundimtare [60]. Theksojmë se përzgjedhja e algorimit grupues mbi matricën e përkatësisë është e pavarur nga përzgjedhja e algoritmit (apo algoritmeve) të përdorur në etapën e gjenerimit të grupimeve. Gjithsesi në modelin tonë në të dy rastet kemi preferuar algoritmin e turbullt të mesoreve, bazuar në shkallën më të mirë të kompleksitetit që karakterizon këtë algoritëm në krahasim me algoritmet e tjerë të grupimit të turbullt.
Modelin e metodës ansambël që projektuam mund ta paraqesim në mënyrë më të përmbledhur me anë të pseudokodit në vijim:

1. Për secilën vlerë të variablit nr nga 0 në $r - 1$ përsërit hapat 2 dhe 3.
2. U jepet vlera fillestare të rastësishme të gjithë qendrave të grupimeve.
3. Zbatohet algoritmi i turbullt i mesoreve mbi bashkësinë e të dhënave duke përdorur vlerat fillestare të pikës 2 me vlera të koeficientit të turbullsisë 1.25, 1.5, 2 dhe 2.5 për të gjeneruar përkatësisht ndarjet U_{4nr+1}, U_{4nr+2}, U_{4nr+3} dhe U_{4nr+3}.
4. Për secilën vlerë të variablit nr nga 0 në $4r - 1$ përsërit hapat 5 , 6 dhe 7.
5. Për secilën vlerë të variablit i nga 1 në n përsërit hapat 6 dhe 7.
6. Për secilën vlerë të variablit j nga 1 në n përsërit hapin 7.
7. Përllogarisim vlerën respektive të matricës së shoqërimit M_{nr}:
$$(M_{nr})_{ij} = \sum_{t=1}^{k} T((U_{nr})_{ti}, (U_{nr})_{tj})$$
8. Përllogarisim vlerat respektive të matricës së përkatësisë M^*:
$$M_{ij}^* = \frac{1}{4r} \sum_{t=0}^{4r-1} T(U_{ti}, U_{tj})$$
9. Aplikojmë algoritmin e turbullt të mesoreve mbi rreshtat e matricës M për të gjeneruar ndarjen përfundimtare në grupime
10. FUND

Figurë 32 - Pseudokodi përshkrues i një modeli homogjen të një metode ansambël të grupimit të turbullt

6.3 Një model heterogjen i një metode ansambël të grupimit të turbullt

Në këtë seksion do të projektojmë një metodë ansambël heterogjene të grupimit të turbullt të cilën do ta implementojmë (në seksionin 6.4) si dhe do të studiojmë eksperimentalisht shkallën e saktësisë së saj. Një metodë ansambël quhet heterogjene kur ndarjet gjenerohen nga algoritme të ndryshme. Algoritmet që do të shfrytëzojmë në këtë metodë për të realizuar etapën e parë të saj (gjenerimi i ndarjeve) janë algoritmi i turbullt i mesoreve, algoritmi Gustafson-Kessel, algoritmi Gath-Geva dhe algoritmi i turbullt i mesoreve i bazuar në bërthama. Do të përzgjidhen disa vlera të ndryshme të qendrave të grupimeve (në mënyrë të rastësishme) dhe për secilën përzgjedhje do të ekzekutohen të katërt algoritmet në fjalë.

Në fazën e dytë (të shkrirjes), ashtu sikurse në modelin homogjen që diskutuam më parë, do të vemë në përdorim përqasjen e shoqërimit. Për secilën ndarje vlerat e matricës së shoqërimit përllogariten me anë të një t-norme, ku ne kemi përzgjedhur t-normën prodhim $T(x, y) = x \cdot y$. Të gjithë matricat e shoqërimit shkrihen më pas në matricën e përkatësisë (e cila përllogaritet si mesatarja aritmetike e të gjithë matricave të shoqërimit). Përfundimisht mbi rreshtat (ose shtyllat) e matricës së përkatësisë zbatojmë përsëri algoritmin e turbullt të mesoreve për të gjeneruar ndarjet përfundimtare.

Modelin e metodës ansambël që projektuam mund ta paraqesim në mënyrë më të përmbledhur me anë të pseudokodit në vijim [60]:

1. Për secilën vlerë të variablit nr nga 0 në $r - 1$ përsërit hapat 2 deri në 6.
2. U jepet vlera fillestare të rastësishme të gjithë qendrave të grupimeve.
3. Zbatohet algoritmi i turbullt i mesoreve mbi bashkësinë e të dhënave duke përdorur vlerat fillestare të pikës 2 për të gjeneruar ndarjen U_{4nr+1}.
4. Zbatohet algoritmi Gustafson-Kessel mbi bashkësinë e të dhënave duke përdorur vlerat fillestare të pikës 2 për të gjeneruar ndarjen U_{4nr+2}.
5. Zbatohet algoritmi Gath-Geva mbi bashkësinë e të dhënave duke përdorur vlerat fillestare të pikës 2 për të gjeneruar ndarjen U_{4nr+3}.
6. Zbatohet algoritmi i turbullt i mesoreve i bazuar në bërthama mbi bashkësinë e të dhënave duke përdorur vlerat fillestare të pikës 2 për të gjeneruar ndarjen U_{4nr+2}.
7. Për secilën vlerë të variablit nr nga 0 në $4r - 1$ përsërit hapat 8 , 9 dhe 10.
8. Për secilën vlerë të variablit i nga 1 në n përsërit hapat 9 dhe 10.
9. Për secilën vlerë të variablit j nga 1 në n përsërit hapin 10.

10. Përllogarisim vlerën respektive të matricës së shoqërimit M_{nr}:

$$(M_{nr})_{ij} = \sum_{t=1}^{k} T((U_{nr})_{ti}, (U_{nr})_{tj})$$

11. Përllogarisim vlerat respektive të matricës së përkatësisë M^*:

$$M_{ij}^* = \frac{1}{4r} \sum_{t=0}^{4r-1} T(U_{ti}, U_{tj})$$

12. Aplikojmë algoritmin e turbullt të mesoreve mbi rreshtat e matricës M^* për të gjeneruar ndarjen përfundimtare në grupime.
13. FUND

Figurë 33- Pseudokodi përshkrues i një modeli heterogjen të një metode ansambël të grupimit të turbullt

6.4 Studime eksperimentale mbi metodat ansambël

Në këtë seksion do të zhvillojmë disa studime eksperimentale mbi dy modelet që paraqitëm (homogjen dhe heterogjen) në lidhje me metodat ansambël në analizën e grupimeve. Për të realizuar procedurat eksperimentale përsëri do të shfrytëzojmë disa bashkësi të dhënash standarde të ofruara publikisht në depozitën e të dhënave të Universitetit të Kalifornisë në Irvine, si dhe një bashkësi të dhënash sintetike. Bashkësitë e të dhënave standarde që kemi përzgjedhur janë Breast Cancer Wisconsin, Dermatology, Vehicle Silhouettes dhe Yeast ndërsa nga bashkësitë sintetike kemi përzgjedhur Synth1. Ndonëse për këto bashkësi struktura e brendshme (grupimet e brendshme) është e njohur paraprakisht, ashtu sikurse në procedurat eksperimentale në kapitullin e katërt, ne do të shfrytëzojmë të gatshëm vetëm numrin e grupimeve. Grupimet e njohura të këtyre bashkësive nuk shfrytëzohen aspak gjatë procesit të analizës së grupimeve me anë të metodave ansambël, por përdoren vetëm në fund për të vlerësuar shkallën e saktësisë së algoritmeve përkatës.

6.4.1 Eksperimente mbi modelin homogjen të metodave ansambël

Modelin homogjen të metodës ansambël të propozuar në seksionit 6.2 e kemi studiuar eksperimentalisht mbi bashkësitë e të dhënave Breast Cancer Wisconsin, Dermatology, Vehicle Silhouettes, Yeast dhe Synth1. Për secilën bashkësi ekperimentet janë kryer disa herë me vlera të parametrit r që varionin përkatësisht në 4, 8, 12 dhe 16, pra vlerat e numrit total të ndarjeve varionin në 16, 32, 48 dhe 64 (kujtojmë se për çdo grup vlerash fillestare algoritmi i turbullt i mesoreve u ekzekutua katër herë duke variuar vlerën e koeficientit të turbullsisë). Më tej me anë të ndarjeve

të përftuara, ndërtuam matricat e shoqërimit, të cilat u kombinuan me njëra-tjetrën për të formuar matricën e përkatësisë, në bazë të së cilës gjenerohet ndarja përfundimtare (siç e përshkruam dhe në pseudokodin e figurës 32). Në fund të eksperimenteve, rezultatet e gjeneruara nga secili eksperiment i krahasojmë me grupimet e njohura të secilës bashkësi për të përllogaritur shkallën e saktësisë. Rezultatet e eksperimenteve i përmbledhim në tabelën e mëposhtme:

Bashkësia e të dhënave	**Algoritmi i turbullt i mesoreve**	**Metoda ansambël homogjene**			
		r = 4	**r = 8**	**r = 12**	**r = 16**
Breast Cancer Wisconsin	0.619	0.629	0.634	0.632	0.636
Dermatology	0.572	0.581	0.587	0.591	0.595
Vehicle Silhouettes	0.570	0.580	0.610	0.610	0.618
Yeast	0.582	0.593	0.621	0.619	0.623
Synth1	0.732	0.751	0.772	0.781	0.785

Tabelë 17 - Shkalla e sakësisë në modelin homogjen të metodave ansambël

Siç shihet dhe në rezultatet e paraqitura në tabelën e mësipërme, metoda ansambël homogjene është pothuajse gjithnjë me shkallë më të lartë saktësie se ekzekutimi i vetëm një algoritmi të grupimit. Një vëzhgim tjetër është se me rritjen e numrit të ndarjeve të gjeneruara (gjatë etapës së parë), përgjithësisht rritet dhe shkalla e saktësisë së algoritmit, por gjithashtu në raste të rralla rritja numrit të mund të mos e përmirësojë shkallën e saktësisë (madje mund ta përkeqësojë lehtas atë). Nëse do të krahasonim shkallën e përmirësimit të metodave ansambël me përqasjen e mësimit gjysmë-të mbikëqyrur, natyrisht kjo e fundit ka një shkallë më të lartë saktësie. Kjo është e pritshme duke qenë se në përqasjen e mësimit gjysmë-të mbikëqyrur kemi të bëjmë me një informacion shtesë të disponueshëm.

6.4.2 Eksperimente mbi modelin heterogjen të metodave ansambël

Modelin heterogjen të metodës ansambël të propozuar në seksionit 6.3 gjithashtu e kemi studiuar eksperimentalisht mbi bashkësitë e të dhënave Breast Cancer Wisconsin, Dermatology, Vehicle Silhouettes, Yeast dhe Synth1. Për secilën bashkësi ekperimentet janë kryer disa herë me vlera të parametrit r që varionin përkatësisht në 4, 8, 12 dhe 16, pra vlerat e numrit total të ndarjeve varionin në 16, 32, 48 dhe 64 (kujtojmë se për çdo grup vlerash fillestare u ekzekutuan katër algoritme,

përkatësisht algoritmi i turbullt i mesoreve, algoritmi Gustafson-Kessel, algoritmi Gath-Geva dhe algoritmi i turbullt i mesoreve i bazuar në bërthama). Më tej me anë të ndarjeve të përftuara, ndërtuam matricat e shoqërimit, të cilat u kombinuan me njëra-tjetrën për të formuar matricën e përkatësisë, në bazë të së cilës gjenerohet ndarja përfundimtare (siç e përshkruam dhe në pseudokodin e figurës 33). Në fund të eksperimenteve, rezultatet e gjeneruara nga secili eksperiment i krahasojmë me grupimet e njohura të secilës bashkësi për të përllogaritur shkallën e saktësisë. Rezultatet e eksperimenteve i përmbledhim në tabelën e mëposhtme:

Bashkësia e të dhënave	**Algoritmi i turbullt i mesoreve**	**Metoda ansambël heterogjene**			
		r = 4	**r = 8**	**r = 12**	**r = 16**
Breast Cancer Wisconsin	0.619	0.630	0.642	0.645	0.660
Dermatology	0.572	0.592	0.597	0.596	0.625
Vehicle Silhouettes	0.570	0.576	0.591	0.613	0.620
Yeast	0.582	0.603	0.617	0.625	0.639
Synth1	0.732	0.760	0.768	0.791	0.805

Tabelë 18 - Shkalla e sakësisë në modelin heterogjen të metodave ansambël

Siç shihet dhe në rezultatet e paraqitura në tabelën e mësipërme, metoda ansambël heterogjene është pothuajse gjithnjë me shkallë më të lartë saktësie se ekzekutimi i vetëm një algoritmi të grupimit. Një vëzhgim tjetër është se me rritjen e numrit të ndarjeve të gjeneruara (gjatë etapës së parë), përgjithësisht rritet dhe shkalla e saktësisë së algoritmit, por gjithashtu në raste të rralla rritja numrit të mund të mos e përmirësojë shkallën e saktësisë (madje mund ta përkeqësojë lehtas atë). Nëse do të krahasonim shkallën e përmirësimit të metodave ansambël heterogjene me përqasjen e mësimit gjysmë-të mbikëqyrur, natyrisht kjo e fundit ka një shkallë më të lartë saktësie. Kjo është e pritshme duke qenë se në përqasjen e mësimit gjysmë-të mbikëqyrur kemi të bëjmë me një informacion shtesë të disponueshëm. Nëse do të krahasonim metodat ansambël homogjene me ato heterogjene, vihet re një avantazh i lehtë i metodave heterogjene ndaj metodave homogjene.

7. PËRFUNDIME, KONTRIBUTE DHE DREJTIME PUNE KËRKIMORE NË TË ARDHMEN

Në këtë kapitull përmbyllës do të përmbledhim përfundimet kryesore të çështjeve të trajtuara gjatë të gjithë këtij punimi, do të vemë në dukje kontributet si rezultat i punës kërkimore (të kryer kryesisht për qëllime krahasuese eksperimentale) si dhe do të skicojmë drejtime në të cilët mund të zhvillohet punë kërkimore e mëtejshme si vazhdë e këtij studimi.

7.1 Përfundime

Logjika e turbullt është një përgjithësim i logjikës klasike, e cila zhvillon konceptin e vërtetësisë së pjesshme (një vlerë vërtetësie në segmentin e vazhduar [0,1]). Ashtu sikurse logjika klasike bazohet në konceptet e bashkësive dhe relacioneve klasike, logjika e turbullt bazohet në bashkësitë dhe relacionet e turbullta (të cilat gjithashtu janë përgjithësime përkatësisht të bashkësive dhe relacioneve klasike). Logjika e turbullt ofron mjaft fleksibilitet në mënyrën si perceptohet, modelohet dhe përpunohet informacioni, duke u bërë një mjet gjerësisht i përdorur në një gamë të gjerë disiplinash.

Analiza e grupimeve është një nga format kryesore të mësimit të pambikëqyrur, qëllimi i së cilës është ndarja e të dhënave në grupime (nënbashkësi) në mënyrë të tillë që të dhënat brenda të njëjtit grupim të jenë sa më të ngjashme me njëra-tjetrën dhe sa më të ndryshme nga të dhënat në grupime të tjera. Aftësia e algoritmeve të grupimeve për të vepruar në kushte të mungesës së informacionit i ka bërë ato një mjet mjaft të rëndësishëm në analizën eksploratore të të dhënave dhe kanë gjetur zbatime në një mori disiplinash si biznes, biologji, mjekësi, përpunim imazhesh, psikologji etj.

Analiza e turbullt grupimeve është një fushë studimi e cila integron koncepte të logjikës së turbullt dhe analizës së grupimeve. Ideja qëndrore është ndërtimi i grupimeve të turbullta ku një element të ketë anëtarësi të pjesshme në disa grupime njëkohësisht. Përqasja e turbullt në problemin e analizës së grupimeve ofron më shumë fleksibilitet dhe natyrshmëri veçanërisht në modelimin e të dhënave nga bota reale. Për realizimin e analizës së turbullt të grupimeve janë zhvilluar një sërë algoritmesh kur ndër më kryesorët janë algoritmi i turbullt i mesoreve, algoritmi Gustafson-Kessel, algoritmi Gath-Geva, algoritmi i turbullt i mesoreve bazuar në bërthama, algoritmi posibilistik i mesoreve etj.

Analiza e turbullt e grupimeve (madje dhe analiza e grupimeve në përgjithësi) duke qenë se vepron në kushte ku nuk ofrohet ndonjë informacion paraprak, përballet me një problem kritik i cili është verifikimi i grupimeve. Për trajtimin e këtij problemi janë zhvilluar një sërë indeksesh vlerësues, të cilët bazohen në parime të përgjithshme si kompaktësia brenda grupimeve dhe ndarja midis grupimeve të ndryshme. Disa nga indekset më të njohur janë koeficienti i particionit, entropia e particionit, indeksi i ndarjes, indeksi Xie-Beni, indeksi Fukuyama-Sugeno, hipervolumet e turbullta etj.

Në disa raste analiza e turbullt grupimeve mund të veprojë në kushte ku jemi në prani të një sasie relativisht të vogël (jo të plotë) informacioni si për shembull disa të dhëna të kategorizuara ose disa kufizime në lidhje me çifte të dhënash për të cilat dihet nëse janë apo jo në të njëjtiun grupim . Në këtë rast nuk jemi më në kushte të mësimit të pambikëqyrur, por mund të veprojmë me përqasje të mësimit gjysmë-të mbikëqyrur. Algoritmi i turbullt i mesoreve mund të përshtatet për të vepruar në kushte të mësimit gjysmë-të mbikëqyrur duke modifikuar funksionin objektiv në mënyrë që të marrë në konsideratë informacionin e diponueshëm.

Algoritmet e analizës së grupimeve karakterizohen nga disa pika të dobëta si ndjeshmëria ndaj vlerave fillestare të grupimeve dhe ndjeshmëria ndaj pranisë së vlerave të largëta apo vlerave me zhurmë. Për të evituar këto disavantazhe mund të përdoren metodat ansambël të analizës së grupimeve. Ideja thelbësore e metodave ansambël është zbatimi disa herë i algoritmeve të grupimit për të gjeneruar një seri ndarjesh mbi bashkësinë e të dhënave, të cilat më pas shkrihen (kombinohen me njëra-tjetrën) për të formuar ndarjen përfundimtare.

7.2 Kontribute

Si kontributet kryesore të këtij punimi konsiderojmë një varg procedurash eksperimentale të kryera me qëllime sintonizimi dhe krahasimi, zbatimin e teknikave të analizës së turbullt mbi të dhënat e studentëve për krijimin e profileve karakterizues të performances, si dhe propozimin e dy modeleve ansambël të analizës së grupimeve dhe krahasimin eksperimental të tyre. Më konkretisht, në vijim po listojmë rezultatet sipas kapitujve.

Në kapitullin e katërt, në pjesën e parë të tij, zhvilluam disa procedura eksperimentale me karakter sintonizues në lidhje me rendin e distancës Minkovski si dhe në lidhje me koeficientin e turbullsisë. Gjatë këtyre eksperimenteve qëllimi ishte jo vetëm analizimi i vlerave të parametrave në fjalë për të cilat algoritmet kishin një shkallë më

të lartë saktësie, por edhe krahasimi i algoritmeve më njëri-tjetrin. Rezultatet e këtyre eksperimenteve treguan për vlerat optimale të rendit të distancës Minkovski midis 1.5 dhe 3, si dhe për vlerat optimale të koeficientit të turbullsisë midis 1.25 dhe 2. Ndërsa në krahasimin e algoritmeve me njëri-tjetrin, vumë re një stabilitet të algoritmit të turbullt të mesoreve të bazuar në bërthama si dhe një luhatshmëri të algoritmit Gath-Geva, i cili konvergjonte më shpejt në optimume lokalë. Përveç kësaj realizuam dhe krahasimin e algoritmeve sipas numrit të iteracionve dhe kohës fizike të ekzekutimit. Në krahasimin sipas numrit të iteracioneve vëzhguam një epërsi të algoritmit klasik të mesoreve si dhe të algoritmit të turbullt të mesoreve, ndërsa në krahasimin sipas kohës fizike të ekzekutimit algoritmi klasik i mesoreve kryeson ndjeshëm, pasuar nga algoritmi i turbullt i mesoreve.

Një varg procedurash eksperimentale të përshkruara gjatë pjesës së dytë të kapitullit të katërt analizojnë problemin e verifikimit të grupimeve. Vëzhgimi kryesor i këtyre eksperimenteve ishte se përgjithësisht zbatimi i një indeksi të vetëm verifikimi nuk garanton zbulimin e grupimeve më cilësore, ndërsa zbatimi i disa indekseve të verifikimit realizon një vlerësim më të qëndrueshëm të grupimeve.

Në pjesën e fundit të kapitullit të katërt kemi propozuar dhe implementuar një metodologji analizuese të të dhënave të studentëve duke ndërtuar profileve karakterizues të performancës së përgjithshme dhe specifike. Metodologjia në fjalë bazohet në disa algoritme të grupimit të turbullt si dhe në shfrytëzimin e indekseve të verifikimit për vlerësimin e grupimeve më cilësore.

Në kapitullin e pestë kemi zhvilluar krahasime eksperimentale të algoritmeve të grupimit të turbullt në kushte të mësimit gjysmë-të mbikëqyrur. Vëzhgimi kryesor i këtyre eksperimenteve ishte se informacioni shtesë e përmirësonte shkallën e saktësisë së algoritmeve të grupimit dhe në shumicën e rasteve me rritjen e sasisë së informacionit rritet dhe shkalla e saktësisë.

Në kapitullin e gjashtë kemi propozuar dhe krahasuar ekperimentalisht dy modele metodash ansambël (një model homogjen dhe një model heterogjen) në trajtimin e problemit të ndjeshmërisë ndaj vlerave fillestare. Këto metoda përmirësojnë qëndrueshmërinë në trajtimin e problemit të analizës së grupimeve.

7.3 Drejtime për zhvillim pune kërkimore në të ardhmen

Ekzistojnë një sërë potencialesh për zhvillim pune kërkimore si vazhdë e këtij studimi dhe në këtë seksion po përmendim disa nga drejtimet kryesore. Së pari në procedurat eksperimentale mund të përfshijmë edhe algoritme të tjerë të analizës së turbullt të grupimit të grupimeve si algoritmi i turbullt i varieteteve, algoritmi i turbullt i

elipsoidëve, algoritmi i turbullt i medoideve, algoritmi i grupimit bazuar në relacione të turbullta etj. Disa nga këto algoritme janë mjaft specifikë dhe përgjithësisisht epërsia e tyre qëndron në bashkësi të dhënash me struktura mjaft të veçanta. Fusha interesante studimi, por më sfiduese në implementim, janë edhe analiza e turbullt e grupimeve e bazuar në logjikën e turbullt intuitive dhe analiza e turbullt e grupimeve bazuar në logjikën e turbullt të llojit 2.

Gjithashtu mund të eksperimentohet në lidhje me modele të analizës së grupimeve të përbërë nga disa faza (disa shtresa). Për shembull algoritmi Gath-Geva i cili karakterizohet nga një konvergjencë e shpejtë në optimume lokale, mund të paraprihet nga një algoritëm tjetër si algoritmi i turbullt i mesoreve. Pra mund të formojmë një model kategorizues i cili të veprojë në fazën e parë me anë të algoritmit të turbullt të mesoreve dhe në fazën algoritmi Gath-Geva të operojë mbi qendrat e përllogaritura gjatë fazës së parë. Ky model mund të krahasohet eksperimentalisht për të analizuar ndikimin në shkallën e saktësisë.

Në studimin eksperimental në lidhje më analizën e turbullt të grupimeve për një përqasje të mësimit gjysmë-të mbikëqyrur mund të përfshijmë një grup më të gjerë algoritmesh, të gjeneruar si rezultat i formave të tjera të modifikimit të funksionit objektiv.

Studimin eksperimental në lidhje me metodat ansambël përveç synimit për trajtimin e problemit të ndjeshmërisë së vlerave fillestare, mund të zhvillohet edhe për trajtimin e ndjeshmërisë ndaj vlerave me zhurmë dhe vlerave të largëta.

Referenca

[1]. Yager, Ronald R., and Lotfi Asker Zadeh, eds. *An introduction to fuzzy logic applications in intelligent systems*. Kluwer Academic, pp 27- 34, 1992.

[2]. Jain, Anil K., M. Narasimha Murty, and Patrick J. Flynn. "Data clustering: a review." *ACM Computing Surveys (CSUR)* 31.3, pp 264-323, 1999.

[3]. Hoppner, F., Klawonn, F. , Kruse, R., and Runkler, T., *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*, pp 12-23 John Wiley & Sons, 1999.

[4]. Abonyi, Janos, and Balazs Feil. *Cluster analysis for data mining and system identification*, pp 8-11, Springer, 2007.

[5]. Xu, Rui, and Don Wunsch. *Clustering*. Vol. 10, pp 83-87, John Wiley & Sons, 2008.

[6]. Zadeh, Lotfi A. "Fuzzy sets." *Information and control* 8.3, pp 338-353, 1965.

[7]. Ross, Timothy J. *Fuzzy logic with engineering applications*, pp 34-40, John Wiley & Sons, 2009

[8]. Klir, G., and Yuan, B. *Fuzzy sets and fuzzy logic* (Vol. 4), pp 33-47, Prentice Hall, 1995

[9]. Pedrycz, Witold, and Fernando Gomide. *An introduction to fuzzy sets:analysis and design*, pp 7-13, MIT Press, 1998.

[10]. Klement, Erich Peter, Radko Mesiar, and Endre Pap. "Triangular norms. Position paper I: basic analytical and algebraic properties." *Fuzzy Sets and Systems* 143.1, 2004

[11]. Butnariu, Dan, Erich Peter Klement, and Samy Zafrany. "On triangular norm-based propositional fuzzy logics." *Fuzzy Sets and Systems* 69.3, pp 241-255, 1995.

[12]. Zimmermann, H. J. *Fuzzy Set Theory and Its Applications Second, Revised Edition*. Kluwer academic publishers, 1992.

[13]. Weber, Siegfried. "A general concept of fuzzy connectives, negations and implications based on t-norms and t-conorms." *Fuzzy sets and systems* 11.1, pp 103-113, 1983.

[14]. Buckley, James J., and Esfandiar Eslami. *An introduction to fuzzy logic and fuzzy sets*. Vol. 13, pp 46-48, Springer, 2002.

[15]. Klir, G., and Yuan, B. *Fuzzy sets and fuzzy logic* (Vol. 4), pp 119-128, Prentice Hall, 1995

[16]. Bandler, Wyllis, and Ladislav J. Kohout. "Special properties, closures and interiors of crisp and fuzzy relations." *Fuzzy sets and Systems* 26.3, pp 317-331, 1988.

[17]. Murali, V. "Fuzzy equivalence relations." *Fuzzy sets and systems* 30.2 (1989): 155-163.

[18]. Baczyński, Michał, and Balasubramaniam Jayaram. *An Introduction to Fuzzy Implications*. Springer Berlin Heidelberg, 2008.

[19]. Fodor, János C. "On fuzzy implication operators." *Fuzzy Sets and Systems* 42.3, 293-300, 1991.

[20]. Burhan Türkşen, I., Vladik Kreinovich, and Ronald R. Yager. "A new class of fuzzy implications. Axioms of fuzzy implication revisited." *Fuzzy sets and systems* 100.1, 267-272, 1998.

[21]. Bedregal, Benjamín C., et al. "On interval fuzzy S-implications." *Information Sciences* 180.8, pp 1373-1389, 2010.
[22]. Atanassov, K. T., and G. Gargov. "Intuitionistic fuzzy-logic.", D*okladi na bolgarskata akademiya na naukite* 43.3, 1990.
[23]. Szmidt, Eulalia. *Distances and similarities in intuitionistic fuzzy sets*. Vol. 307 pp 8-19, Springer, 2014.
[24]. Vlachos, Ioannis K., and George D. Sergiadis. "Intuitionistic fuzzy information–applications to pattern recognition." *Pattern Recognition Letters* 28.2 197-206, 2007.
[25]. Castillo, Oscar, and Patricia Melin. *3 Type-2 Fuzzy Logic*, pp 29-35, Springer Berlin Heidelberg, 2008.
[26]. Liang, Qilian, and Jerry M. Mendel. "Interval type-2 fuzzy logic systems: theory and design." *Fuzzy Systems, IEEE Transactions on* 8.5, 535-550, 2000.
[27]. Bedalli, Erind "Fuzzy logic applications in modeling and estimation of economical quantities", ISTI-2012
[28]. Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." *Neural Networks, IEEE Transactions on* 16.3, 645-678, 2005.
[29]. Yang, M-S. "A survey of fuzzy clustering." *Mathematical and Computer modeling*, 18.11, 1993.
[30]. Von Luxburg, Ulrike. *Statistical learning with similarity and dissimilarity functions*, pp 9-16 Logos-Verlag, 2004.
[31]. Jain, Anil. "Data clustering: 50 years beyond K-means." *Pattern Recognition Letters* 31.8, 651-666, 2010
[32]. Bedalli, Erind "An overview of fuzzy hierarchical clustering methods", ISTI-2013
[33]. Miyamoto, Sadaaki, Hidetomo Ichihashi, and Katsuhiro Honda. *Algorithms for fuzzy clustering: methods in c-means clustering with applications*. Vol. 229, 99-107, Springer, 2008.
[34]. Ding, Chris, and Xiaofeng He. "Cluster merging and splitting in hierarchical clustering algorithms." *Proceedings IEEE International Conference*, 2002.
[35]. Bedalli, Erind and Ninka, Ilia "A classification model combining the fuzzy K-NN and fuzzy C-Means algorithms", IWBCMS-2013
[36]. Han, Jiawei, and Micheline Kamber. *Data Mining, Southeast Asia Edition: Concepts and Techniques*, pp 450-455, Morgan kaufmann, 2006.
[37]. Bedalli, Erind dhe Ninka, Ilia "Një studim krahasues mbi disa variante të algoritmit të grupimit të turbullt të mesoreve", Buletini Shkencor i Universitetit "Aleksander Xhuvani" Elbasan (Edicioni i Shkencave të Natyrës), 2013
[38]. Gustafson, Donald E., and William C. Kessel. "Fuzzy clustering with a fuzzy covariance matrix." *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*. Vol. 17. IEEE, 1978.
[39]. Liu, Hsiang-Chuan, et al. "Fuzzy C-means algorithm based on standard mahalanobis distances." *Proceedings of the International Symposium on Information Processing*, 2009.
[40]. Gath, Isak, and Amir B. Geva. "Unsupervised optimal fuzzy clustering." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 11.7, 773-780, 1989

[41]. Grabocka, Josif, Erind Bedalli, and Lars Schmidt-Thieme. "Efficient Classification of Long Time-Series." *ICT Innovations 2012*, pp 47-57, Springer Berlin Heidelberg, 2013.
[42]. Graves, Daniel, and Witold Pedrycz. "Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study." *Fuzzy sets and systems* 161.4, 522-543, 2010
[43]. Krishnapuram, Raghuram, and James M. Keller. "The possibilistic c-means algorithm: insights and recommendations." *Fuzzy Systems, IEEE Transactions on* 4.3, pp 385-393, 1996.
[44]. Blake, Catherine, and Christopher J. Merz. "{UCI} Repository of machine learning databases.", 1998.
[45]. Halkidi, Maria, Yannis Batistakis, and Michalis Vazirgiannis. "On clustering validation techniques." *Journal of Intelligent Information Systems* 17.2-3, pp 107-145., 2001.
[46]. Bedalli, E., Ninka, I., *Implementation of some cluster validity methods for fuzzy cluster analysis*, ISCIM 2013
[47]. Wang, Weina, and Yunjie Zhang. "On fuzzy cluster validity indices." *Fuzzy sets and systems* 158.19, pp 2095-2117, 2007.
[48]. Bedalli, E., Ninka, I., "Exploring an educational system's data through fuzzy cluster analysis ", 11th Annual International Conference on Information Technology & Computer Science, Athens 2014
[49]. Grabocka, J., Bedalli E., and Schmidt-Thieme L. "Supervised Nonlinear Factorizations Excel In Semi-supervised Regression." *Advances in Knowledge Discovery and Data Mining*, pp188-199, Springer International Publishing, 2014.
[50]. E. Bair, "Semi-supervised clustering methods.", Wiley Interdisciplinary Reviews, pp 349-361, 2013.
[51]. A. Bouchachia, W. Pedrycz. "Data clustering with partial supervision.", Data Mining and Knowledge Discovery, 45-49, 2006.
[52]. D.T.C. Lai, J.M. Garibaldi, "A comparison of distance-based semi-supervised fuzzy c-means clustering algorithms", Fuzzy Systems (FUZZ), IEEE International Conference 2011.
[53]. Bedalli, E., Ninka, I., *Adapting the fuzzy c-means clustering algorithm for a semi-supervised learning approach*, International Journal of Science, Innovation and New Technology (IJSINT), vol 1/2014
[54]. Strehl, Alexander, and Joydeep Ghosh. *Cluster ensembles-a knowledge reuse framework for combining multiple partitions*. The Journal of Machine Learning Research, pp 583-617, 2003.
[55]. De Oliveira, J. Valente, and W. Pedrycz, eds. *Advances in fuzzy clustering and its applications*. New York: Wiley, pp 69-83, 2007
[56]. Strehl, Alexander, and Joydeep Ghosh. *Cluster ensembles-a knowledge reuse framework for combining partitionings.* AAAI/IAAI. 2002.
[57]. Hadjitodorov, Stefan T., Ludmila I. Kuncheva, and Ludmila P. Todorova. "Moderate diversity for better cluster ensembles." Information Fusion 7.3, pp 264-275, 2006.

[58]. Kuncheva, L. I., S. T. Hadjitodorov, and L. P. Todorova. *Experimental comparison of cluster ensemble methods.* Information Fusion, 9th International Conference on. IEEE, 2006

[59]. Avogadri, Roberto, and Giorgio Valentini. *Ensemble clustering with a fuzzy approach, Supervised and Unsupervised Ensemble Methods and their Applications*. Springer Berlin Heidelberg, 2008.

[60]. Bedalli, Erind, and Ilia Ninka. "Using homogeneous fuzzy cluster ensembles to address fuzzy c-means initialization drawbacks.", International Journal of Scientific and Engineering Research (IJSER), vol.6, June 2014