UO‘K 004.85

YANGI YADROLARNING AJRALUVCHANLIK INDEKSI YORDAMIDA GRAFLARGA ASOSLANGAN KLASTER YADROLARINING AJRALISHINI TAHLIL QILISH

Adilova F.T.[1], [+] *Davronov R.R.*[1]

[1] O‘zbekiston Respublikasi Fanlar akademiyasi V.I. Romanovskiy nomidagi Matematika instituti, Toshkent, O‘zbekiston

[+] rifqat@gmail.com

Annotatsiya. Ushbu maqolada graflarga asoslangan zichlik klasterlash algoritmlarining sifatini yaxshilash va avtomatlashtirish uchun mo‘ljallangan Yadrolarning ajraluvchanlik indeksi (Core Separation Index, CSI) deb nomlangan klasterlashning yangi ichki validlik indeksi taqdim etilgan. Taklif etilgan indeks har bir yadroning ichki bog‘liqligi (cohesion) va uning boshqa yadrolar bilan bog‘liqligi (coupling) o‘rtasidagi nisbatni o‘lchash orqali klaster yadrolarini ajratish sifatini baholaydi. Biz CSI indeksining nazariy asosini, uning xossalari haqidagi teoremani va qat’iy isbotini keltiramiz. Indeks graf zichligi variatsiyasini tahlil qilishga asoslangan (Clustering by graph density variation analysis, GDVA) algoritmiga integratsiya qilingan bo‘lib, klasterlashning optimal parametrlarini avtomatik ravishda aniqlash imkonini beradi. CSI indeksini DVI(density-based cluster validity indices), Silhouette va Dunn kabi mavjud indekslar bilan har tomonlama eksperimental taqqoslash turli tuzilishdagi beshta sintetik ma’lumotlar to‘plamida amalga oshirildi. Adjusted Rand Index (ARI) va Normalized Mutual Information (NMI) tashqi metrikalari yordamida baholangan natijalar shuni ko‘rsatadiki, CSI eng yaxshi klasterlash sifatini ta’minlaydi (o‘rtacha NMI 0.7460), hisoblash tezligi bo‘yicha boshqa indekslardan sezilarli darajada ustun (o‘rtacha 0.17s). Taklif etilayotgan indeksning tashqi sifat ko‘rsatkichlari bilan yuqori korrelyatsion bog‘liqligi ham ko‘rsatilgan bo‘lib, bu uning klasterlarning optimal tuzilmasini baholash va tanlash uchun ishonchli ekanligini tasdiqlaydi.

Kalit so‘zlar: klaster tahlili, graflar asosida klasterlash, zichlikli klasterlash, klasterlarning validlik indeksi, GDVA, CSI, parametrlarni avtomatik aniqlash, ARI, NMI.

1 KIRISH

Klaster tahlili, ma’lumotlarni intellektual tahlil qilish (Data Mining) va mashinali o‘rganishning asosiy vazifalaridan biri bo‘lib, ma’lumotlar to‘plamini guruhlarga (klasterlarga) shunday ajratishga qaratilganki, bunda bitta klaster ichidagi obyektlar bir-biriga maksimal darajada o‘xshash, turli klasterlardagi obyektlar esa bir-biridan maksimal darajada farq qiladi [1]. So‘nggi yillarda ma’lumotlarni graf ko‘rinishida modellashtiradigan graflar asosida klasterlash usullariga katta e’tibor qaratilmoqda, bunda cho‘qqilar obyektlarga mos keladi va qirralarning og‘irliklari ularning o‘xshashlik darajasini aks ettiradi [2]. Bunday yondashuvlar, ayniqsa, k-means kabi an’anaviy usullarning zaif tomoni bo‘lgan ixtiyoriy shakl va tuzilishdagi klasterlarni aniqlashda samarali hisoblanadi [3]. Graf usullar orasida DBSCAN [4] va uning ko‘plab modifikatsiyalari kabi zichlikka asoslangan algoritmlar alohida o‘rin tutadi.

Ushbu usullar klasterlarni zichligi past bo‘lgan hududlar bilan ajratilgan zich hududlar sifatida aniqlaydi. Ushbu sohadagi istiqbolli yo‘nalishlardan biri [5] da taklif etilgan graf zichligi variatsiyasi tahlili (Graph Density Variation Analysis, GDVA) hisoblanadi. GDVA usuli har bir klaster zich “yadro”ga ega degan taxminga asoslangan bo‘lib, klasterlarni aniqlash va keyinchalik ushbu yadrolarni kengaytirish orqali aniqlash imkonini beradi. Biroq, boshqa ko‘plab klasterlash algoritmlari singari, GDVA giperparametrlar to‘plamiga bog‘liq bo‘lib, ularni qo‘lda tanlash murakkab va ko‘p vaqt talab qiladigan masala bo‘lib, ekspert bilimlari va ko‘plab tajribalarni talab qiladi.

Natijaviy bo‘linishning sifati bu parametrlarning to‘g‘ri tanlanishiga bevosita bog‘liq. Ushbu muammoni hal qilish uchun klasterlashning validlik indekslari (Cluster Validity Indices, CVI) qo‘llaniladi, ular olingan bo‘linish sifatini miqdoriy baholash imkonini beradi [6]. CVI sinflarning haqiqiy belgilari mavjudligini talab qiladigan tashqi va faqat ma’lumotlarning o‘zi va klasterlarning tuzilishi asosida sifatni baholaydigan ichki turlarga bo‘linadi. Silhouette [7], Dunn [8] va Davies-Bouldin [9] kabi mavjud ichki indekslar sferik shakldagi klasterlar uchun o‘zini yaxshi ko‘rsatgan, ammo ko‘pincha zichlik usullarining

kuchli tomoni bo‘lgan murakkab shakldagi tuzilmalar uchun samarasiz. Bundan tashqari, indekslarni hisoblash ko‘p resurs talab qilishi mumkin, bu esa parametrlarni avtomatik sozlashning iterativ protseduralarida ulardan foydalanishni qiyinlashtiradi.

Ushbu ishda biz yangi ichki validlik indeksini - yadrolarning ajralish indeksini taqdim etamiz. Ushbu indeks graflarga asoslangan zichlik algoritmlarida klaster yadrolarini ajratish sifatini baholash uchun maxsus ishlab chiqilgan. CSI har bir yadroning ichki bog‘liqligi va uning qo‘shni yadrolar bilan bog‘liqligi muvozanatiga asoslanib, yadrolarning bir-biridan qanchalik yaxshi ajratilganligini baholaydi.

Ushbu maqoladagi bizning asosiy hissamiz quyidagilardan iborat: 1) CSI yangi validlik indeksini shakllantirish va nazariy asoslash; 2) Optimal delta parametrini avtomatik aniqlash uchun CSI indeksini GDVA algoritmiga integratsiyalash; 3) Turli ma’lumotlar to‘plamlarida ma’lum indekslar (DVI, Silhouette, Dunn) bilan CSI kompleks qiyosiy tahlilini o‘tkazish; 4) CSI dan foydalanish nafaqat ARI va NMI metrikalari bo‘yicha klasterlash sifatini oshiradi, balki optimal parametrlarni topish uchun zarur bo‘lgan vaqtni sezilarli darajada qisqartiradi.

Maqola quyidagi tuzilishga ega. 2-bo‘limda mavzu bo‘yicha adabiyotlar sharhi keltirilgan. 3-bo‘limda CSI indeksining rasmiy ta’rifi va uning nazariy asoslari keltirilgan. 4-bo‘limda tajribalarning modifikatsiyalangan algoritmi va metodologiyasi keltirilgan. 5-bo‘limda tajriba natijalari keltirilgan va tahlil qilingan.

2 ADABIYOTLAR TAHLILI

Graflar asosida klasterlash muammosi so‘nggi o‘n yilliklarda faol o‘rganilmoqda. Ushbu bo‘limda biz ushbu sohadagi asosiy ishlarni ko‘rib chiqamiz, ayniqsa zichlikka asoslangan usullar va validlik indekslariga e’tibor qaratamiz.

2.1 Graflar asosida klasterlash usullari

Graf usullarini shartli ravishda bir nechta toifalarga bo‘lish mumkin. Spektral klasterlash [10] ma’lumotlarni kichikroq o‘lchamli fazoga proyeksiyalash uchun Laplas grafi matritsasining xususiy vektorlaridan foydalanadi, bu yerda klasterlar chiziqli ajraladigan bo‘ladi. Normallashtirilgan kesimlar usuli (Normalized Cut) [11] eng mashhur misollardan biri bo‘lib, tasvirlarni segmentlashda keng qo‘llaniladi, ammo u klasterlar sonini oldindan belgilashni talab qiladi va taxminan bir xil o‘lchamdagi klasterlarni shakllantirishga moyil. Single-linkage algoritmi [12] kabi ierarxik usullar klasterlarning ichki tuzilishini ifodalovchi dendrogrammani quradi. Ushbu usullar klasterlar sonini berishni talab qilmaydi, lekin shovqinga sezgir (“zanjir” effekti) va ularning hisoblash murakkabligi katta ma’lumotlar to‘plamlari uchun yuqori. Markov klaster algoritmi (MCL) [1] kabi oqimlar va tasodifiy adashishlarga asoslangan usullar zich hududlarni aniqlash uchun grafdagi oqimlarni modellashtiradi. MCL oqsil tarmoqlarini klasterlash uchun bioinformatikada o‘zini yaxshi ko‘rsatdi, ammo uning samaradorligi zich graflarda pasayadi. Affinity Propagation [14] algoritmi bir vaqtning o‘zida klasterlarning markazlarini (“namunalar”) topadi va ularga “xabarlar” almashinuvi orqali qolgan nuqtalarni taqsimlaydi, ammo u ham yuqori hisoblash murakkabligiga ega.

2.2 Zichlikka asoslangan usullar

Zichlikka asoslangan algoritmlar graf usullarining kichik sinfi bo‘lib, klasterlarni past zichlikdagi sohalar bilan ajratilgan yuqori zichlikdagi sohalar sifatida belgilaydi. DBSCAN [4] klassik algoritmi ushbu yo‘nalishning tamal toshi hisoblanadi. U ixtiyoriy shakldagi klasterlarni topishga qodir va shovqinga chidamli. Biroq, uning unumdorligi o‘zgaruvchan zichlikdagi ma’lumotlar uchun tanlash qiyin bo‘lgan ikkita parametrga (eps va MinPts) bog‘liq. Ushbu muammoni hal qilish uchun ko‘plab kengaytmalar taklif qilindi. OPTICS [15] turli zichlikdagi klasterlarni ajratib olishga imkon beradigan tartiblangan ma’lumotlar tuzilmasini quradi, ammo algoritmning o‘zi aniq bo‘linishni ko‘rsatmaydi. HDBSCAN [16] algoritmi ushbu g‘oyaning zamonaviy rivojlanishi bo‘lib, klasterlar iyerarxiyasini tuzadi va ulardan eng barqarorlarini avtomatik ravishda tanlaydi, bu esa uni bugungi kunda eng kuchli va mashhur usullardan biriga aylantiradi. Thang Le [5] dissertatsiyasida tavsiflangan GDVA (Graph Density Variation Analysis) usuli boshqacha yondashuvni taklif qiladi. U grafning “kuchsiz” cho‘qqilarini (eng kichik lokal zichlik bilan) iterativ olib tashlash va bu jarayonning dinamikasini tahlil qilishga asoslangan. Zichlikning keskin pasayishi klasterlar yadrolariga tegishli cho‘qqilarning olib tashlanishi haqida signal beradi deb taxmin qilinadi. Ushbu usul bizning ishimizda qo‘llaniladigan GDVA algoritmiga asos bo‘ldi. GDVAning afzalligi uning hisoblash samaradorligi va ma’lumotlarning ierarxik tuzilishini aniqlash qobiliyatidir.

2.3 Zichlikka asoslangan usullar

Maqsadli o‘zgaruvchilar mavjud bo‘lmaganda klasterlash sifatini baholash muhim vazifa hisoblanadi. Ichki validlik indekslari klasterlar ichidagi ixchamlikni va ular o‘rtasidagi bo‘linishni baholaydi. Har bir nuqta uchun Silhouette indeksi [7] uning qo‘shni klasterga qaraganda o‘z klasteriga qanchalik yaqinligini baholaydi. U qavariq klasterlar uchun yaxshi ishlaydi, ammo uning qiymatlari murakkab shaklli klasterlar uchun noinformativ bo‘lishi mumkin. Dunn indeksi [8] minimal klasterlararo masofaning maksimal klaster ichki diametriga nisbati sifatida aniqlanadi. U shovqin va chekka qiymatlatga sezgir, chunki u ekstremal qiymatlardan foydalanadi. Devis-Boldin indeksi [1] har bir klaster uchun unga eng “o‘xshash” boshqa klasterni topadi va bu qiymatlarni o‘rtalashtiradi. Silhouette singari, u sferik klasterlar uchun ko‘proq mos keladi.

Zichlik usullari uchun maxsus boshqa yondashuvlar ishlab chiqilgan. GDVA bilan bir xil ishda taklif qilingan DVI (Density-Based Validity Indices) [5,18] indekslari har bir nuqta uchun klaster ichidagi va klasterlararo zichlik o‘rtasidagi farqni yoki nisbatni baholaydi. Bu ularni zichlik klasterlash natijalarini baholash uchun yanada moslashtiradi. So‘nggi tadqiqotlarda [17] ma’lumotlarga kichik o'zgartirishlar kiritilganda bo‘linish qanchalik o‘zgarishini baholovchi klaster barqarorligiga asoslangan indekslar taklif etilgan. Mavjud indekslarning xilma-xilligiga qaramay, murakkab ma’lumotlar tuzilmalari uchun klasterlash sifati bilan yaxshi bog‘liq bo‘lgan tezkor va ishonchli ko‘rsatkichlarga bo‘lgan ehtiyoj hali ham mavjud bo‘lib, bu bizning CSI indeksini ishlab chiqishga turtki bo‘ldi.

3 YADROLARNING AJRALUVCHANLIK INDEKSI

3.1 Ta’riflar

Faraz qilaylik, $G=(V,E,W)$ vaznli yo’naltirilmagan graf bo‘lsin, bu yerda V - uchlar to‘plami, E - qirralar to‘plami, W esa- vaznlar matritsasi, unda w_{uv} qiymati u va v uchlar orasidagi o‘xshashlikni ifodalaydi. Faraz qilaylik, klasterlash algoritmi klasterlar yadrolari to‘plamini ajratib oldi: $C_{cores}=\{S_1,S_2,\ldots,S_k\}$, bu yerda har bir $S_i \subseteq V$ yadrosi i -klasterning zich markaziy qismiga tegishli bo‘lgan cho‘qqilarning qism to‘plami hisoblanadi.

1-ta’rif: Yadroning ichki bog‘langanligi (Cohesion). S_i yadrosining ichki bog‘liqligi $Coh(S_i)$ deb belgilanadi, bu yadro ichidagi qirralarning o‘rtacha og‘irligi sifatida aniqlanadi:

$$Coh(S_i)=\frac{\sum_{u,v\in s_j,u\neq v} w_{uv}}{|E_i|},$$

bu yerda $|E_i|$ - S_j uchlar to‘plami tomonidan induksiyalangan qism grafdagi qirralar soni. Agar $|E_i|=0$ bo‘lsa, u holda $Coh(S_i)=0$ bo‘ladi. Bu metrika bitta yadro ichidagi cho‘qqilarning qanchalik zich va kuchli bog‘langanligini aks ettiradi.

2-ta’rif: Yadrolarning o’zaro tutashligi (Coupling). Ikki xil S_i va $S_j\,(i\neq j)$ yadrolari orasidagi o‘zaro tutashligi, $Coup(S_i,S_j)$ deb belgilanuvchi, S_i uchlarini S_j uchlari bilan bog‘laydigan qirralarning o‘rtacha og‘irligi sifatida aniqlanadi:

$$Coup(S_i,S_j)=\frac{\sum_{u\in S_i,v\in S_j} w_{uv}}{|S_i|\cdot|S_j|}.$$

Bu metrika ikki xil yadroning o‘rtacha tutashligini ko‘rsatadi. Intuitiv ravishda, yaxshi yadrolarga bo‘linish har bir yadroning yuqori ichki bog‘liqligi va turli yadrolar o‘rtasidagi past tutashligi bilan tavsiflanishi kerak.

3-ta’rif: Yadrolarning ajraluvchanlik indeksi. Berilgan yadrolar to‘plami $\{S_1,\ldots,S_k\}$ uchun CSI har bir yadro bog‘liqligining uning boshqa har qanday yadro bilan maksimal tutashligiga nisbatining o‘rtacha qiymati sifatida aniqlanadi:

$$CSI=\frac{1}{k}\sum_{i=1}^{k}\frac{Coh(S_i)}{max_{j\neq i}\{Coup(S_i,S_j)+\epsilon\}},$$

bu yerda ϵ - yadrolar to‘liq izolyatsiyalangan bo‘lsa, nolga bo‘linishning oldini olish uchun kichik konstanta (masalan, 10^{-9}).

CSI ning yuqori qiymati yadrolar ichida zich va bir-biri bilan kuchsiz bog‘langanligini ko‘rsatadi, bu klaster yadrolarining sifatli ajralishiga mos keladi.

3.2. Teorema va isbot

1-teorema. Grafning $\{S_1,\ldots,S_k\}$ yadrolarga ajratilishi berilgan bo‘lsin. Agar ikkita eng eng kuchli bog‘langan S_a va S_b yadrolarni yangi $S_{new}=S_a\cup S_b$ yadroga birlashtirsak, yangi $Coh(S_{new})$ yadrosining ichki bog‘liqligi dastlabki yadrolarning o‘rtacha o‘lchangan bog‘liqligidan oshmasa va S_a bilan S_b o‘rtasidagi tutashlig ustunlik qilsa, u holda yangi $\{S_1,\ldots,S_{new},\ldots S_k\}$ yadrolar to‘plami uchun CSI qiymati oshmaydi (bu yerda S_a va S_b, S_{new} bilan almashtirilgan).

Isbot. CSI qiymatining o‘zgarishini ko‘rib chiqamiz. Boshlang‘ich qiymat:

$$CSI_{old}=\frac{1}{k}\left(\frac{Coh(S_a)}{M_a}+\frac{Coh(S_b)}{M_b}+\sum_{i\neq a,b}\frac{Coh(S_i)}{M_i}\right),$$

bu yerda $M_i=max_{j\neq i}\left\{Coup\left(S_i,S_j\right)\right\}$. Aytaylik, S_a va S_b ikkita yadro bo‘lib, ular uchun $Coup(S_a,S_b)$ barcha juft tutashligilar orasida maksimal bo‘lsin. U holda $M_a\approx Coup(S_a,S_b)$ va $M_b\approx Coup(S_a,S_b)$.

Birlashtirishdan keyin yangi CSI qiymati:

$$CSI_{new}=\frac{1}{k-1}\left(\frac{Coh(S_{new})}{M_{new}}+\sum_{i\neq a,b}\frac{Coh(S_i)}{M_i^{'}}\right),$$

bunda $S_{new}=S_a\cup S_b$.

Yangi $Coh\left(S_{new}\right)$ yadrosining ichki bog‘liqligi S_a ichidagi, S_b ichidagi va S_a va S_b o‘rtasidagi bog‘lanishlarni o‘z ichiga oladi. U boshlang‘ich kattaliklar orqali ifodalanishi mumkin:

$$Coh\left(S_{new}\right)=\frac{|E_a|Coh(S_a)+|E_b|Coh(S_b)+|S_a||S_b|Coup(S_a,S_b)}{|E_a|+|E_b|+|S_a||S_b|}.$$

Agar $Coh\left(S_{new}\right)$ o‘rtacha o‘lchangan $Coh(S_a)$ va $Coh(S_b)$ dan yuqori emas deb taxmin qilinsa, bu $Coup(S_a,S_b)$ ning hissasi anomal darajada katta emasligini anglatadi.

Yangi yadro uchun maksimal tutashligi $M_{new}=max_{j\neq a,b}\left\{Coup\left(S_{new},S_j\right)\right\}$.

S_a va S_b eng tutash bo‘lganligi sababli, M_{new} ni $Coup(S_a,S_b)$ dan kichik deb taxmin qilish oqilona bo‘ladi.

CSI_{new}dagi $\sum_{i\neq a,b}\frac{Coh(S_i)}{M_i^{'}}$ hadlar CSI_{old} dagi mos hadlarga juda yaqin bo‘ladi, chunki ularning maksimal tutashligi o‘zgarmaydi.

Asosiy o‘zgarish CSI_{old} ning birinchi ikkita hadida va CSI_{new} ning birinchi hadida sodir bo‘ladi

$$\frac{Coh(S_a)}{M_a}+\frac{Coh(S_b)}{M_b}\approx\frac{Coh(S_a)+Coh(S_b)}{Coup(S_a,S_b)}.$$

Yangi had esa $\frac{Coh\left(S_{new}\right)}{M_{new}}$ ga teng.

M_a va M_b katta (S_a va S_b o‘rtasidagi kuchli bog‘liqlik tufayli) va M_{new} kichik bo‘lganligi sababli, yangi haddagi maxraj kamaydi. Biroq $Coh\left(S_{new}\right)$ surati ham o‘zgardi.

Agar birlashma xato bo‘lgan bo‘lsa (ya’ni S_a va S_b aslida turli klasterlarga tegishli bo‘lsa), u holda $Coup(S_a,S_b)$ $Coh(S_a)$ va $Coh(S_b)$ ga nisbatan kichik bo‘lgan. Birlashish $Coh\left(S_{new}\right)$ ning past

bo‘lishiga olib keladi (chunki kuchsiz yadrolararo aloqalar bilan “kuchsizlanadi”), bu esa umumiy CSI qiymatining pasayishiga olib keladi.

Aksincha, agar S_a va S_b bitta haqiqiy yadroning qismlari bo‘lsa, u holda $Coup(S_a, S_b)$ ularning ichki bog‘liqligi bilan taqqoslanadi. Bu holda $Coh(S_{new})$ yuqori bo‘ladi va CSI qiymati oshishi mumkin.

Teorema shuni ta’kidlaydiki, eng tutash yadrolar noto‘g‘ri birlashtirilganda (bu ko‘plab algoritmlarda “ochko‘z” qadam hisoblanadi), CSI qiymati oshmaydi. Bu xususiyat CSI ni maksimal darajada oshirish kerak bo‘lgan baholash funksiyasi uchun yaxshi nomzodga aylantiradi.Yadrolarni iterativ ravishda birlashtiradigan algoritm CSI qiymati o‘sishdan to‘xtaganda to‘xtashi mumkin.

4 ALGORITM VA METODOLOGIYA

4.1. CSI yordamida GDVA algoritmi

Biz CSI indeksini GDVA usuliga asoslangan GDVA algoritmiga kiritdik. Asosiy g‘oya yadrolarni ajratib olish chegarasini nazorat qiluvchi delta parametrining optimal qiymatini avtomatik ravishda topish uchun CSI dan foydalanishdir.

Algoritm quyidagi qadamlardan iborat:

- Grafni qurish: kirish ma’lumotlar to‘plami uchun vaznli qo‘shnilik grafi quriladi.
- Zichlik variatsiyasi tahlili: zichligi eng kichik bo‘lgan cho‘qqilarni iterativ olib tashlash protsedurasi bajariladi, natijada D_t (t qadamdagi minimal zichlik) va M_t (t qadamdagi uzoqlashgan cho‘qqi) ketma-ketliklari quriladi.
- CSI yordamida optimal deltani qidirish: a. Deltani qidirish uchun diapazon beriladi (masalan, 0.1 dan 0.9 gacha). b. Diapazondan har bir delta qiymati uchun: i. D_t va M_t asosida yadrolarga nomzod cho‘qqilar to‘plami ajratiladi. ii. Ushbu nomzod cho‘qqilar $\{S_1, \dots, S_k\}$ yadrolari bo‘lgan bog‘liq komponentlarga bo‘linadi. iii. CSI ($\{S_1, \dots, S_k\}$) qiymati hisoblanadi. v. CSI maksimal bo‘lgan delta qiymati tanlanadi.
- Yakuniy klasterlash: Optimal delta yordamida klasterlarning yakuniy yadrolari aniqlanadi.
- Yadrolarning kengayishi: Qolgan cho‘qqilar (yadrolarga kirmaydigan) eng yaqin klasterlarga taqsimlanib, yakuniy bo‘linishni hosil qiladi.

4.2. Metodologiya

Taklif etilgan yondashuvning samaradorligini baholash uchun biz CSI ni boshqa indekslar: DVI1, DVI2, DVI3, Silhouette va Dunn bilan taqqoslaydigan bir qator tajribalar o‘tkazdik.

Ma’lumotlar to‘plami. Scikit-learn yordamida hosil qilingan 5 ta sintetik ma’lumotlar to‘plamidan foydalanildi: Iris, 3 ta yaxshi ifodalangan klasterli klassik ma’lumotlar to‘plami; Blobs: Aniq ajratilgan 4 ta sferik klasterli ma’lumotlar; Blobs_Hard: 5 ta kesishuvchi sferik klasterli ma’lumotlar; Moons: Yarim oy shaklidagi ikkita chiziqli bo‘lmagan klaster, Circles: Bir-biriga kiritilgan ikkita klaster.

Taqqoslanadigan indekslar.Optimal deltani topish uchun maqsad funksiyasi sifatida csi, dvi1, dvi2, dvi3, silhouette, dunn lardan navbatma-navbat foydalanildi.

Baholash metrikalari. Yakuniy klasterlash sifati ikkita tashqi ko‘rsatkich yordamida baholandi: Adjusted Rand Index (ARI): tasodifiy mosliklarni hisobga olgan holda ikkita bo‘linmaning o‘xshashligini o‘lchaydi.Normalized Mutual Information (NMI): Entropiyaga nisbatan normallashtirilgan haqiqiy va olingan bo‘linishlar o‘rtasidagi o‘zaro ma’lumotlarni o‘lchaydi.

Protsedura: har bir ma’lumotlar to‘plami va har bir ichki indeks uchun [0.1, 0.9] oralig‘ida optimal deltani qidirish protsedurasi ishga tushirildi. Keyin topilgan delta bilan to‘liq klasterlash amalga oshirildi va natija ARI va NMI bo‘yicha baholandi. Shuningdek, optimal deltani topish uchun sarflangan vaqt ham o‘lchandi.

5 TAJRIBALAR NATIJALARI VA MUHOKAMASI

5.1. Klasterlash sifatini taqqoslash

1-jadvalda optimal delta parametrini topish uchun turli xil ichki indekslardan foydalangan holda olingan ARI va NMI o‘rtacha qiymatlari keltirilgan.

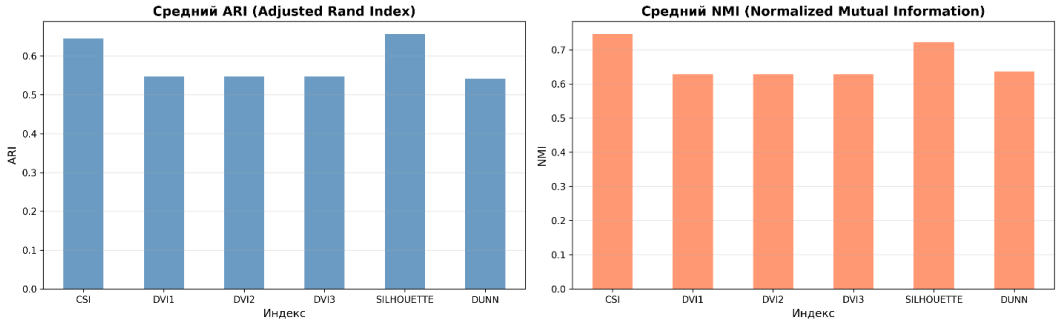
Jadvaldan ko‘rinib turibdiki, maqsad funksiyasi sifatida CSI indeksidan foydalanish NMI metrikasi bo‘yicha eng yaxshi o‘rtacha klasterlash sifatiga erishish imkonini beradi (0,7460). ARI metrikasi bo‘yicha eng yaxshi natijani Silhouette (0.6564) ko‘rsatadi, ammo CSI undan biroz ortda qoladi (0.6451). Shuni

ta’kidlash kerakki, DVI indeksining barcha uchta varianti bir xil va pastroq natijalarni ko‘rsatdi. Dunn indeksi eng yomon sifatni ko‘rsatdi.

1-jadval. Turli validlik indekslari uchun o‘rtacha sifat ko‘rsatkichlari (ARI va NMI) va bajarish vaqti

Indeks	O‘rtacha ARI	O‘rtacha NMI	O‘rtacha vaqt (s)
CSI	0.6451	**0.7460**	**0.17**
DVI1	0.5480	0.6287	0.25
DVI2	0.5480	0.6287	0.25
DVI3	0.5480	0.6287	0.25
SILHOUETTE	**0.6564**	0.7226	0.22
DUNN	0.5418	0.6365	0.81

CSI ning asosiy afzalligi uning ishlash tezligidir. CSI yordamida optimal parametrni topishning o‘rtacha vaqti atigi 0,17 soniyani tashkil etdi, bu barcha raqobatchilarga qaraganda ancha tez. Eng sekin ko‘rsatkich Dunn indeksi (0.81 s) bo‘ldi. Bu CSIni katta ma’lumotlar to‘plami bilan ishlash uchun ayniqsa jozibador qiladi, bu yerda validlik indeksini qayta-qayta hisoblash tor joyga aylanishi mumkin. Ko‘rgazmalilik uchun ushbu natijalar 1- va 2-rasmlarda keltirilgan.



1-rasm. Turli indekslar uchun ARI va NMI ko‘rsatkichlari bo‘yicha o‘rtacha klasterlash sifatini taqqoslash

2-rasm. Turli indekslar uchun optimal parametrlarni topish uchun sarflangan o‘rtacha vaqtni taqqoslash

5.2. Tashqi ko‘rsatkichlar bilan bog‘liqlik

Ichki validlik indeksining ishonchliligini uning tashqi ko‘rsatkichlar bilan bog‘liqlik darajasi orqali baholash mumkin. Yuqori korrelyatsiya ichki indeks klasterlashning "haqiqiy" sifatini yaxshi bashorat qilishini anglatadi. 2-jadvalda ichki indekslar qiymatlari va ARI/NMI yakuniy baholari o‘rtasidagi Pirson va Spirmen korrelyatsiya koeffitsiyentlari keltirilgan.

Korrelyatsiya tahlili shuni ko‘rsatadiki, Silhouette indeksi ikkala tashqi ko‘rsatkichlar, ayniqsa NMI bilan eng yuqori korrelyatsiyaga ega (Pirson: 0.879, Spirmen: 0.900). Bu shuni anglatadiki, Silhouette qiymatini yuqori ehtimollik bilan maksimal darajada oshirish klasterlashning yaxshi sifatiga olib keladi.

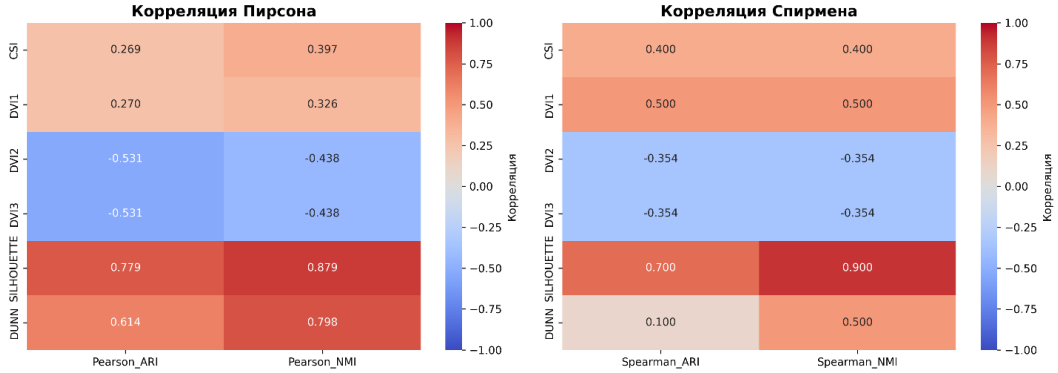
Dunn indeksi ham Pirson bo‘yicha NMI bilan kuchli korrelyatsiyani ko‘rsatadi (0,798), ammo Spirmen bo‘yicha zaif.

2-jadval. Ichki indekslar va tashqi ko‘rsatkichlar o‘rtasidagi korrelyatsiya

Indeks	Pearson-ARI	Pearson-NMI	Spearman-ARI	Spearman-NMI
CSI	0.269	0.397	0.400	0.400
DVI1	0.270	0.326	0.500	0.500
DVI2	-0.531	-0.438	-0.354	-0.354
DVI3	-0.531	-0.438	-0.354	-0.354
SILHOUETTE	**0.779**	**0.879**	**0.700**	**0.900**
DUNN	0.614	0.798	0.100	0.500

CSI va DVI1 indekslari kuchsiz ijobiy korrelyatsiyani ko‘rsatadi, DVI2 va DVI3 esa o‘rtacha salbiy korrelyatsiyani ko‘rsatadi (bu kutilganidek, chunki ularni minimallashtirish kerak).

Korrelyatsiyalarni vizuallashtirish 3-rasmda keltirilgan.



3-rasm. Ichki indekslar va tashqi metrikalar o‘rtasidagi Pirson va Spirmen korrelyatsiyalarining issiqlik xaritasi (ARI, NMI)

5.3. Muhokama

Olingan natijalar bir nechta muhim xulosalar chiqarish imkonini beradi. Birinchidan, taklif etilgan CSI indeksi klasterlash sifati va hisoblash tezligi o‘rtasidagi eng yaxshi muvozanatni ko‘rsatadi. Rasmiy korrelyatsiya ko‘rsatkichlari bo‘yicha Silhouette indeksidan past bo‘lishiga qaramay, amalda GDVA algoritmi uchun maqsad funksiyasi sifatida qo‘llanilishi NMIning eng yuqori o‘rtacha qiymatiga olib keldi. Bu CSI yadrolarni ajratishga asoslangan algoritmning o‘ziga xos xususiyatlariga ko‘proq mos kelishi bilan izohlanishi mumkin. Silhouette universal indeks sifatida yakuniy sifat bilan yaxshi korrelyatsiyalanadi, ammo uni maksimallashtirish har doim ham aynan yadroviy yondashuv uchun optimal bo‘lgan parametrlarni tanlashga olib kelmaydi. Ikkinchidan, CSI ning hisoblash samaradorligi uning shubhasiz afzalligidir. U boshqa ko‘rib chiqilgan indekslarga nisbatan 1,5-5 marta tezroq ishlaydi. Bu xususiyat parametrlarni uzoq vaqt tanlamasdan tezkor avtomatik klasterlashni talab qiladigan masalalar uchun juda muhimdir.

Silhouette bilan solishtirganda CSI ning tashqi metrikalar bilan zaif korrelyatsiyasi CSI butun bo‘linishni emas, balki faqat yadro sifatini baholashi bilan bog‘liq bo‘lishi mumkin. Yakuniy sifat ikkinchi bosqich - yadrolarning kengayishiga ham bog‘liq. Ehtimol, ba’zi hollarda yaxshi ajratilgan yadrolar optimal yakuniy bo‘linishga olib kelmaydi. Bu kelajakdagi tadqiqotlar uchun yo‘nalish ochadi: yadrolarning sifatini ham, ularning kengayish sifatini ham hisobga oladigan tarkibiy indeksni ishlab chiqish.

Shunga qaramay, CSI ni maksimallashtirishga asoslangan strategiya amalda juda yaxshi natijalar berishi (eng yaxshi NMI) yadrolarni sifatli ajratish muvaffaqiyatli zichlik klasterlash uchun asosiy qadam ekanligini va CSI bu qadamni boshqarish uchun samarali vosita ekanligini ko‘rsatadi.

6 XULOSA

Ushbu maqolada zichlikka asoslangan graf algoritmlari uchun mo‘ljallangan yangi ichki CSI klasterlash validlik indeksi (Yadro bo‘linish indeksi) taklif etildi. Biz uning rasmiy ta’rifini, nazariy asosini keltirib, amaliy samaradorligini ko‘rsatib berdik.

Tajribalar shuni ko‘rsatdiki, GDVA algoritmiga CSI integratsiyasi DVI, Silhouette va Dunn kabi boshqa mashhur indekslarni NMI metrikasi bo‘yicha ortda qoldirib, klasterlashning yuqori sifatiga olib

keladigan parametrlarni avtomatik ravishda topish imkonini beradi. CSI ning asosiy afzalligi uning yuqori hisoblash tezligi bo‘lib, bu uni avtomatik klasterlash masalalarida maqsad funksiyasi sifatida ishlatish uchun ideal nomzodga aylantiradi.

Kelajakdagi tadqiqotlar yanada kengroq real ma’lumotlar asosida CSI xususiyatlarini o‘rganishga, shuningdek, yadro sifatini baholash va yakuniy bo‘linishni birlashtirgan gibrid indekslarni ishlab chiqishga qaratilgan bo‘lishi mumkin.

ADABIYOTLAR

[1] *Jain, A.K.* "Data clustering: 50 years beyond K-means." Pattern recognition letters 31.8 (2010): 651-666.
[2] *Schaeffer, S.E.* "Graph clustering." Computer science review 1.1 (2007): 27-64.
[3] *Von Luxburg, U.* "A tutorial on spectral clustering." Statistics and computing 17.4 (2007): 395-416.
[4] *Ester, M., et al.* "A density-based algorithm for discovering clusters in large spatial databases with noise." Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996.
[5] *Le, T.V.* "Clustering by graph density variation analysis (GDVA) with density-based cluster validity indices (DVI)." Dissertation, Rutgers University, 2011.
[6] *Arbelaitz, O., et al.* "A comprehensive survey of cluster validity indices." Pattern Recognition 46.1 (2013): 243-258.
[7] *Rousseeuw, P.J.* "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." Journal of computational and applied mathematics 20 (1987): 53-65.
[8] *Dunn, J.C.* "A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters." Journal of cybernetics 3.3 (1973): 32-57.
[9] *Davies, D.L., & Bouldin, D.W.* "A cluster separation measure." IEEE transactions on pattern analysis and machine intelligence 1 (1979): 224-227.
[10] *Ng, A.Y., Jordan, M.I., & Weiss, Y.* "On spectral clustering: Analysis and an algorithm." Advances in neural information processing systems 14 (2001).
[11] *Shi, J., & Malik, J.* "Normalized cuts and image segmentation." IEEE Transactions on pattern Analysis and machine intelligence 22.8 (2000): 888-905.
[12] *Gower, J.C., & Ross, G.J.* "Minimum spanning trees and single linkage cluster analysis." Journal of the Royal Statistical Society: Series C (Applied Statistics) 18.1 (1969): 54-64.
[13] *Van Dongen, S.* "Graph clustering by flow simulation." PhD thesis, University of Utrecht, 2000.
[14] *Frey, B.J., & Dueck, D.* "Clustering by passing messages between data points." science 315.5814 (2007): 972-976.
[15] *Ankerst, M., et al.* "OPTICS: ordering points to identify the clustering structure." ACM Sigmod record. Vol. 28. No. 2. 1999.
[16] *Campello, R.J., Moulavi, D., & Sander, J.* "Density-based clustering based on hierarchical density estimates." Pacific-Asia conference on knowledge discovery and data mining. Springer, Berlin, Heidelberg, 2013.
[17] *Vendramin, L., Campello, R. J., & Hruschka, E.R.* "On the comparison of relative clustering validity criteria." Proceedings of the SIAM International Conference on Data Mining. 2009.
[18] *Давронов, Р.* (2025). Графовый алгоритм кластеризации на основе вариации плотности. Международный Журнал Теоретических и Прикладных Вопросов Цифровых Технологий, 8(2), 58–64. https://doi.org/10.62132/ijdt.v8i2.264.

Поступила в редакцию 29.07.2025

Citation: *Adilova F.T., Davronov R.R.* (2025). Yangi yadrolarning ajraluvchanlik indeksi yordamida graflarga asoslangan klaster yadrolarining ajralishini tahlil qilish. Raqamli texnologiyalarning nazariy va amaliy masalalari xalqaro jurnali. 8(4). –B. 129-137. https://doi.org/10.62132/ijdt.v8i4.313.

ANALYSIS OF GRAPH-BASED CLUSTER CORE SEPARATION USING A NEW CORE SEPARABILITY INDEX

Adilova F.T.[1], [+] *Davronov R.R.*[1]

[1] V.I. Romanovsky Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan Republic of Uzbekistan, Tashkent, Uzbekistan

[+] rifqat@gmail.com

Abstract. This article presents a new internal validity index of clustering called the Core Separation Index (CSI), designed to improve the quality and automate graph-based density clustering algorithms. The proposed index assesses the quality of cluster nucleus separation by measuring the ratio between the internal cohesion of each nucleus and its coupling with other nuclei. We present the theoretical basis of the CSI index, the theorem about its properties, and a strict proof. The index is integrated into the algorithm based on graph density variation analysis (GDVA), which allows for the automatic determination of optimal clustering parameters. A comprehensive experimental comparison of the CSI index with existing indices, such as DVI (density-based cluster validity indices), Silhouette, and Dunn, was carried out on five sets of synthetic data of different structures. The results, assessed using the external metrics Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), show that CSI provides the best clustering quality (average NMI 0.7460), significantly exceeding other indices in calculation speed (average 0.17s). A high correlation of the proposed index with external quality indicators is also shown, which confirms its reliability for assessing and selecting the optimal structure of clusters.

Keywords: cluster analysis, graph-based clustering, density clustering, cluster validity index, GDVA, CSI, automatic parameter determination, ARI, NMI.

АНАЛИЗ РАЗДЕЛИМОСТИ ЯДЕР КЛАСТЕРОВ НА ОСНОВЕ ГРАФОВ С ИСПОЛЬЗОВАНИЕМ НОВОГО ИНДЕКСА CSI

Адилова Ф.Т.[1], [+] *Давронов Р.Р.*[1]

[1] Институт математики имени В.И.Романовского академии наук Республики Узбекистан, Ташкент, Узбекистан

Аннотация. В данной статье представлен новый внутренний индекс валидности кластеризации, названный Индексом Разделимости Ядер (Core Separation Index, CSI), предназначенный для улучшения качества и автоматизации плотностных алгоритмов кластеризации на основе графов. Предложенный индекс оценивает качество выделения ядер кластеров путем измерения соотношения между внутренней связностью (cohesion) каждого ядра и его сопряженностью (coupling) с другими ядрами. Мы представляем теоретическое обоснование индекса CSI, включая теорему о его свойствах и строгое доказательство. Индекс интегрирован в алгоритм GDVA, основанный на анализе вариации плотности графа, что позволяет автоматически определять оптимальные параметры кластеризации. Проведено всестороннее экспериментальное сравнение индекса CSI с существующими индексами, такими как DVI, Silhouette и Dunn, на пяти синтетических наборах данных различной структуры. Результаты, оцененные с помощью внешних метрик Adjusted Rand Index (ARI) и Normalized Mutual Information (NMI), показывают, что CSI обеспечивает лучшее качество кластеризации (в среднем NMI 0.7460) и значительно превосходит другие индексы по скорости вычислений (в среднем 0.17с). Также продемонстрирована высокая корреляция предложенного индекса с внешними метриками качества, что подтверждает его надежность для оценки и выбора оптимальной структуры кластеров.

Ключевые слова: кластерный анализ, кластеризация на основе графов, плотностная кластеризация, индекс валидности кластеров, GDVA, CSI, автоматическое определение параметров, ARI, NMI.