Модуль IY.

№25 дәріс: Кластерлік талдау және оның биологияда қолданылуы.

Мақсаты. Кластерлік талдау әдісінің биологиядағы қолданылуын зерттеу, оның артықшылықтары мен мүмкіндіктерін көрсету.

Күтілетін нәтиже. Студент кластерлік талдаудың мақсатын және негізгі ұғымдарын және дендрограмманы оқып, кластерлердің қалыптасуын түсіндіреді, кластерлік талдауды биологиялық және экологиялық деректерге қолдану жолдарын ұсынады.

Әдістері: мультимедиялық сүйемелдеумен дәріс.

Дәрістің топтама-тәсімі (тірек конспектісі немесе тезистер).

1. Кластерлік талдаудың мәні, мақсаты және негізгі ұғымдары.

2. Кластерлік талдаудың негізгі әдістері және оны жүргізу кезеңдері

3. Негізгі компоненттер әдісі: көпөлшемді деректерді ықшамдау және визуализациялау.

4. Кластерлік талдау мен негізгі компоненттер әдісінің биология және экологияда қолданылуы.

1. Кластерлік талдаудың мәні, мақсаты және негізгі ұғымдары.

Кластерлік талдау – зерттелетін нысандарды олардың белгілері бойынша салыстырып, өзара ұқсас нысандарды бір топқа, яғни кластерге біріктіретін көпөлшемді статистикалық әдістер жиынтығы. Кластер ішіндегі нысандар бір-біріне барынша ұқсас, ал әртүрлі кластерлерге кіретін нысандар бір-бірінен барынша ерекшеленуі қажет.

Кластерлік талдаудың негізгі мақсаты – алдын ала белгіленген топтар болмаған жағдайда деректердің табиғи құрылымын анықтау. Сондықтан ол бақылаусыз оқыту әдістерінің қатарына жатады. Кластерлеуде зерттеу нысандары даралар, популяциялар, түрлер, өсімдік қауымдастықтары, су айдындары, гендер немесе биологиялық сынамалар болуы мүмкін.Кластерлік талдау объектілерді олардың айырмашылығына немесе ұқсастығына қарай санатқа топтастыру үшін қолданылатын бірнеше көп өлшемді (көп вариантты) әдістерді қамтиды. Алгоритм-бір жағынан топ ішіндегі вариацияларды азайту және оларды екінші жағынан топтар арасында көбейту.

Кластерлік талдау пәні (ағылш. "кластер" - кластер, шоқ, топ) 1939 жылы психолог Роберт Трион тұжырымдады. Кластерлік талдаудың классиктері - американдық таксономистер Роберт Сокел және Питер Снит. Кластерлік талдаудың көмегімен мәліметтер массивін бөлу (кластерлеу, кластерлеу) мәселесі шешіледі, осылайша барлық объект сен, бір кластерге (класқа, топқа) тиесілі, басқа кластерлердің объектілеріне қарағанда бір-біріне ұқсас.

Ұқсастықты анықтау үшін Евклид пен Манхэттен арақашықтығы, Пирсонның корреляция коэффициенті, χ2 қашықтығы, Жаккар мен Съеренсен ұқсастық коэффициенттері, Ренконен және басқалары сияқты әр түрлі қашықтық өлшемдері қолданылады.

Көбінесе кластерлер арасындағы қашықтықты анықтайтын иерархиялық алгоритмдер қолданылады (талданатын мәліметтер массивіндегі топтар). Нәтижелер барлық жіктелетін объектілердің бағыныштылығы (иерархиясы) туралы ақпаратты ұсынатын дендрограммалар түрінде графикалық түрде ұсынылады.

Иерархиялық әдістер. Кластерлік талдауын жүзеге асыру үшін бірнеше негізгі алгоритмдер қолданылады, олардың ішінде ең көп тарағандары көрші (жалғыз байланыс) әдісі, алыстағы көрші (толық байланыс) әдісі, орташа байланыс әдісі және Вард әдісі.

2. Кластерлік талдаудың негізгі әдістері және оны жүргізу кезеңдері

Кластерлік талдау әдістері иерархиялық және иерархиялық емес болып бөлінеді. Иерархиялық кластерлеу кезінде нысандардың біртіндеп бірігуі немесе бөлінуі нәтижесінде кластерлер жүйесі қалыптасады. Оның нәтижесі дендрограмма арқылы көрсетіледі.

Агломеративтік тәсілде әр нысан бастапқыда жеке кластер ретінде қарастырылып, кейін ең ұқсас кластерлер біртіндеп біріктіріледі.

Кластерлерді біріктірудің кең таралған тәсілдері:

- ең жақын көрші әдісі;
- ең алыс көрші әдісі;
- топтар арасындағы орташа байланыс әдісі;
- Уорд әдісі.

Иерархиялық емес кластерлеуде кластер саны алдын ала беріледі. Оның кең таралған түрі – k-орташа әдісі. Бұл әдіс нысандарды берілген k кластерге бөліп, әр нысанды өзіне жақын кластер орталығына орналастырады.

Кластерлік талдаудың негізгі кезеңдері:

1. зерттеу мақсаты мен бақылау бірліктерін анықтау;
2. кластерлеуге енгізілетін белгілерді таңдау;
3. жетіспейтін және қате мәндерді тексеру;
4. әртүрлі өлшем бірліктеріндегі көрсеткіштерді стандарттау;
5. арақашықтық немесе ұқсастық өлшемін таңдау;
6. кластерлеу алгоритмін қолдану;
7. кластерлердің оңтайлы санын анықтау;
8. нәтижені дендрограмма немесе кластерлік график арқылы көрсету;
9. кластерлерге биологиялық түсініктеме беру.

Мысалы, көлдерді температурасы, pH, тұздылығы, нитрат мөлшері және түрлік алуантүрлілігі бойынша кластерлерге бөлуге болады.3. Негізгі компоненттер әдісі: көпөлшемді деректерді ықшамдау және визуализациялау.

3. Негізгі компоненттер әдісі: көпөлшемді деректерді ықшамдау және визуализациялау.

Негізгі компоненттер әдісі, яғни Principal Component Analysis – PCA, өзара байланысқан көптеген сандық айнымалыларды ақпараттың басым бөлігін сақтай отырып, аз санды жаңа компоненттерге түрлендіреді. Әрбір негізгі компонент бастапқы айнымалылардың сызықтық комбинациясы болып табылады.

Бірінші негізгі компонент деректердегі өзгергіштіктің ең үлкен бөлігін түсіндіреді. Екінші компонент бірінші компонентке тәуелсіз болады және қалған өзгергіштіктің мүмкіндігінше көп бөлігін қамтиды. Нәтижесінде ондаған белгіні екі немесе үш негізгі компонент арқылы көрнекі түрде көрсетуге болады.

Негізгі компоненттер әдісінің міндеттері:

- көпөлшемді деректер санын азайту;
- өзара корреляцияланған белгілерді біріктіру;
- нысандар арасындағы ұқсастықты визуализациялау;
- маңызды айнымалыларды анықтау;
- шектен тыс немесе ерекше нысандарды табу;
- кейінгі кластерлік талдауды жеңілдету.

PCA мен кластерлік талдаудың мақсаттары бірдей емес. PCA айнымалылар санын қысқартып, деректердегі өзгергіштіктің негізгі бағыттарын көрсетеді. Кластерлік талдау нысандарды ұқсастығы бойынша топтарға бөледі. Зерттеу барысында алдымен PCA жүргізіп, кейін негізгі компоненттердің мәндері бойынша кластерлеу қолданылуы мүмкін.

Мысалы, өсімдіктердің 15 морфометриялық белгісі зерттелсе, PCA бұл белгілерді «өсімдік мөлшері», «жапырақ пішіні» және «генеративтік өнімділік» тәрізді бірнеше жинақталған компонент арқылы сипаттауға мүмкіндік береді.

4. Кластерлік талдау мен негізгі компоненттер әдісінің биология және экологияда қолданылуы.

Кластерлік талдау биологияның әртүрлі салаларында күрделі деректерді топтастыру үшін қолданылады. Таксономияда түрлер мен популяциялар морфологиялық, анатомиялық немесе генетикалық белгілері бойынша салыстырылады. Генетикада гендердің экспрессиялық белсенділігі ұқсас гендер бір кластерге біріктіріледі. Микробиологияда микроағзалар биохимиялық және молекулалық белгілеріне қарай топтастырылады.

Экологияда кластерлік талдау:

- өсімдік қауымдастықтарын жіктеуге;
- зерттеу учаскелерін түрлік құрамы бойынша топтастыруға;
- су айдындарын гидрохимиялық көрсеткіштері бойынша салыстыруға;
- ластану деңгейі ұқсас аумақтарды анықтауға;
- популяциялардың экологиялық және морфологиялық айырмашылықтарын бағалауға;
- биоалуантүрліліктің кеңістіктік заңдылықтарын зерттеуге қолданылады.

Мысалы, 20 зерттеу учаскесіндегі өсімдік түрлерінің кездесуі мен проективтік жамылғысы тіркелсе, кластерлік талдау флоралық құрамы ұқсас учаскелерді біріктіреді. Нәтижесінде шалғындық, далалық, шөлейттік немесе антропогендік өзгеріске ұшыраған қауымдастықтар ажыратылуы мүмкін.

Негізгі компоненттер әдісі осы учаскелердің айырмашылығын анықтайтын басты факторларды көрсетеді. Бірінші компонент топырақ ылғалдылығы мен өсімдік жамылғысына, екінші компонент тұздылық пен антропогендік жүктемеге байланысты болуы ықтимал. Кластерлік талдау топтарды анықтаса, PCA олардың қандай көрсеткіштер бойынша ерекшеленетінін түсіндіруге көмектеседі.

Бақылау сұрақтары:

Кластерлік талдаудың негізгі мақсаты қандай?

Кластер мен айнымалы ұғымдарының айырмашылығы неде?

Иерархиялық және иерархиялық емес кластерлеу қалай ерекшеленеді?

Дендрограмма қандай ақпарат береді?

Деректерді кластерлеу алдында не үшін стандарттау қажет?

Әдебиеттер:

1) Аубакиров Х.Ә. Биометрия: оқулық / Х.Ә.Аубакиров. – А.: Дәуір, 2011. – 188-197 б.

2) Раманқұлова А.А. Биологиялық статистика: оқу құралы. – Алматы: Эверо, 2023. – 95-100 бет.