Изкуствен интелект - летен семестър, 2023/2024 учебна година

Тема 12:
Клъстерен анализ. Основни методи за нейерархична и йерархична клъстеризация

Процесът на „разбиване“ на дадено множество от физически или абстрактни обекти на групи от *сходни* обекти се нарича *клъстеризация*. Един *клъстер* представлява колекция от обекти, които приличат един на друг вътре в клъстера и се различават от обектите от други клъстери.

Клъстеризацията на данни е област с непрекъснато развитие. Интензивни изследвания в тази област се провеждат в такива научни дисциплини като ИЗД, математическа статистика, машинно самообучение, разпознаване на образи, управление на мултимедийни бази от данни, биоинформатика, астрономия, маркетинг и др.

Като клон на *математическата статистика* изследванията върху клъстерния анализ от много години се фокусират основно върху *базиран на разстояние (distance-based)* подход към тази задача. Инструментите за клъстерен анализ, използващи класически, базирани на разстояние алгоритми за клъстеризация като *k*-means, *k*-medoids и др., са част от големи пакети за статистически анализ, като S-Plus, SPSS и др. В *машинното самообучение* клъстеризацията е пример на неуправлявано (без учител – unsupervised) обучение.

В отличие от класификацията, клъстеризацията и самообучението без учител не използват предварително определени класове и вече класифицирани обучаващи примери. По тази причина клъстеризацията може да се разглежда по-скоро като форма на *самообучение чрез наблюдения*, отколкото като самообучение чрез примери. При *концептуалната клъстеризация* една група от обекти формира клъстер само ако той може да бъде описан като понятие. Този подход се отличава от традиционната клъстеризация, която измерва сходството на базата на геометричното разстояние. Концептуалната клъстеризация се извършва на два етапа: (1) откриване на подходящи клъстери и (2) формиране на описание за всеки клъстер, както при класификацията. И в този случай се прилагат критериите за силно вътре-клъстерно и слабо между-клъстерно сходство.

Клъстеризацията е предмет на интензивни научни изследвания, при които нейните потенциални приложения налагат свои специфични изисквания. Методите за клъстеризация, които се разработват и се прилагат в ИЗД, трябва да удовлетворяват посочените по-долу изисквания.

- *Разширяемост.* Съществуват много алгоритми за клъстеризация, които работят много добре върху малки множества от данни, съдържащи не повече от 200 обекта. Но една голяма база от данни може да съдържа милиони обекти. Клъстеризацията, извършена върху определена *извадка* от тези данни, може да доведе до изкривени резултати. Необходими са алгоритми за клъстеризация, които могат да работят върху много големи множества от данни.

- *Възможност за работа с различни типове атрибути.* Съществуват множество алгоритми, предназначени/подходящи за клъстеризация на непрекъснати (числови) данни. Някои приложения обаче може да изискват клъстеризация на данни от други типове, например двоични, номинални или от смесени типове.
- *Откриване на клъстери с различна форма.* Повечето от клъстеризиращите алгоритми се базират върху Евклидовото или абсолютното разстояние. Такива алгоритми имат тенденция да намират сферични по форма клъстери с близък размер и плътност (гъстота). Един клъстер обаче може да има произволна форма и поради това е важно да бъдат разработвани алгоритми, способни да намират клъстери с произволна форма.

- *Минимизиране на изискванията към знанията за проблемната област, необходими за определяне на входни(те) параметри.* Много от клъстеризиращите алгоритми изискват от потребителя да въвежда определени параметри за желания клъстерен анализ (например броя на клъстерите). Резултатите от клъстеризацията могат в значителна степен да зависят от стойностите на тези параметри. Определянето на подобни параметри често е доста трудна задача, особено за множества от данни, съдържащи обекти с голяма размерност. Това не само затруднява потребителите, но и води до проблеми с контрола върху качеството на резултата от клъстеризацията.

- *Възможност за работа със зашумени данни.* За повечето от реалните бази от данни са факт наличието на грешни или липсващи стойности. Някои от клъстеризиращите алгоритми са чувствителни към подобни данни, което води до намаляване на качеството на резултата от клъстеризацията.
- *Нечувствителност към реда на постъпване на входните данни.* Някои клъстеризиращи алгоритми са чувствителни към подредбата на входните данни – например едно и също множество от данни, когато е предоставено на такъв алгоритъм, при различна подредба на данните може да бъде „разбито“ на напълно различни клъстери. По тази причина е важно да бъдат разработвани алгоритми, нечувствителни към подредбата на данните.

- *Голяма размерност.* Една база от данни може да съдържа голямо количество атрибути. Много клъстеризиращи алгоритми са добри при работа с данни с малка размерност, например с двумерни или тримерни данни. Човешките очи са много добър инструмент за определяне на качеството на клъстеризацията в пространство с до 3 размерности. Предизвикателството е да се клъстеризират обекти в многомерни пространства, особено като се отчита, че подобни данни могат да бъдат силно разпръснати и да имат много асиметрично разпределение в съответното пространство.

- *Клъстеризация, базирана на ограничения.* Някои реални приложения може да изискват извършване на клъстеризация при наличието на различни видове ограничения. Да предположим, че задачата е да бъдат избрани в даден град места за разполагане на даден брой банкомати. За решаването на тази задача жилищните сгради може да се клъстеризират с отчитане на специфични ограничения като наличие на минаващи през града реки, пътища и др. В общия случай задачата за намиране на групи в масиви от данни с добро клъстеризационно поведение, които удовлетворяват зададени ограничения, е сериозна изследователска задача.

- *Разбираемост и използваемост.* Потребителите очакват резултатите от клъстеризацията да се поддават на интерпретация, да бъдат разбираеми и използваеми. С други думи, често клъстеризацията трябва да бъде свързана със семантичната интерпретация и очакваните приложения. Важно е да бъде разбрано как една приложна цел може да повлияе на избора на конкретен метод за клъстеризация.

Категоризация на методите за клъстеризация

Съществуващите методи за клъстеризаця могат да бъдат групирани в няколко категории.

- *Методи за разделяне (partitioning)*. При зададена база от данни, съдържаща *n* обекта, един метод за разделяне „разбива“ данните на *k* групи – клъстери, като $k \leq n$. Всички групи заедно трябва да удовлетворяват следните изисквания: (1) всяка група трябва да съдържа най-малко един обект и (2) всеки обект трябва да принадлежи на точно една група.

Последното изискване може да бъде отслабено при някои методи за *размито разделяне*. При зададения брой групи (клъстери), подлежащи на конструиране, методът за разделяне създава някакво първоначално „разбиване“. След това той прилага избраната конкретна *итеративна техника за преместване* на обекти, опитвайки се да подобри качеството на „разбиване“ чрез преместване на обекти от една група в друга. Най-общ критерий за качеството на разделянето е „близостта“ на обекти, намиращи се в един и същ клъстер, и „отдалечеността“ на обекти, намиращи се в различни клъстери. Съществуват различни други критерии за оценка на качеството на разделянето.

За да бъде получено едно глобално оптимално решение при клъстеризация чрез разделяне, е необходимо изчерпващо претърсване на всички възможни „разбивания“. Вместо това в повечето случаи се използва един от следните два популярни евристични метода: (1) алгоритъм *k*-means, при който всеки клъстер се представя чрез средното на всички обекти от клъстера, и (2) алгоритъм *k*-medoids, при който клъстерът се представя чрез един от обектите, разположени близо до центъра на клъстера. Тези евристични методи за клъстеризация дават много добри резултати при намиране на сферични по форма клъстери в бази от данни с малки или средни размери. За намирането на клъстери със сложна форма или за клъстеризация на много големи бази от данни методите, базирани на разделяне, се нуждаят от разширение.

- *Йерархични методи*. Един метод от този вид създава йерархична декомпозиция на даденото множество от обекти (данни). Йерархичните методи за клъстеризация могат да бъдат разделени на *агломеративни (обединяващи)* и *разделящи (devisive)* в зависимост от начина за формиране на йерархичната декомпозиция. *Агломеративният подход* (който често се нарича подход *отдолу нагоре*) започва работата си, третирайки всеки обект като отделна група – клъстер. След това той последователно слива групите, които са „близки“ помежду си, докато или всички групи се слеят в една (най-горното ниво в йерархията), или бъде удовлетворен зададеният критерий за прекратяване на сливането.

Разделящият подход (известен още като подход *отгоре надолу*) започва работата си с включване на всички обекти в един клъстер. След това той итеративно „разбива“ всеки клъстер на по-малки клъстери, докато или всеки обект бъде поместен в отделен клъстер, или бъде удовлетворен зададеният критерий за прекратяване на разделянето.

- *Методи, базирани на плътност.* Повечето методи за „разбиване“ се базират на изчисляване на разстояние между обектите. Такива методи могат да намират само сферични по форма клъстери и срещат големи затруднения при намирането на клъстери с произволна форма. Съществуват други клъстеризиращи методи, базиращи се на понятието *плътност.* Основната идея е да се продължава разрастването на дадения клъстер до тогава, докато неговата плътност (т.е. броят обекти или точки от данни в него) остава над определен праг. С други думи, за всеки обект, намиращ се в дадения клъстер, неговата околност със зададен радиус трябва да съдържа най-малко определен минимум от обекти. Подобен метод може да се използва за филтриране на шум (крайности), както и за откриване на клъстери с произволна форма.

- *Решетъчни методи*. Решетъчните методи „дискретизират“ цялото пространство от обекти на крайно множество от клетки, които формират съответна решетъчна структура. Всички операции по клъстеризацията се изпълняват върху тази структура (т.е. върху дискретизираното пространство). Основното предимство на този подход е бързото време за обработка, което обикновено не зависи от броя на обектите, а само от броя на клетките във всяка от размерностите на това дискретизирано пространство.

- *Методи, базирани на модели.* Методите, базирани на модели, предполагат, че всеки клъстер може да бъде описан с помощта на определен модел, и опитват да „нагласят“ по най-добрия начин наличните данни към тези модели. Един алгоритъм, базиран на модели, може да локализира клъстерните центрове чрез конструиране на определена функция за вероятностната плътност, описваща пространственото разпределение на обектите. Този подход позволява автоматично да се намира броят на клъстерите, базирайки се на стандартни статистики, както и да се определят шумове и екстремни обекти.

Клъстеризация чрез разделяне (partitioning)

При зададени база от данни, съдържаща описания на *n* обекта, и желания брой *k* на клъстерите, в които трябва да бъдат групирани тези данни, един разделящ клъстерен алгоритъм „разбива“ данните на *k* непресичащи се раздела *(k ≤ n)*, всеки от които представя един клъстер. Клъстерите се формират така, че да оптимизират определен критерий за разделяне, наричан често *функция на сходство*, така че обектите от един клъстер са „сходни“, докато обектите от различни клъстери са „различни“ в термините на атрибутите от базата от данни.

Използване на центроиди: метод *k*-means

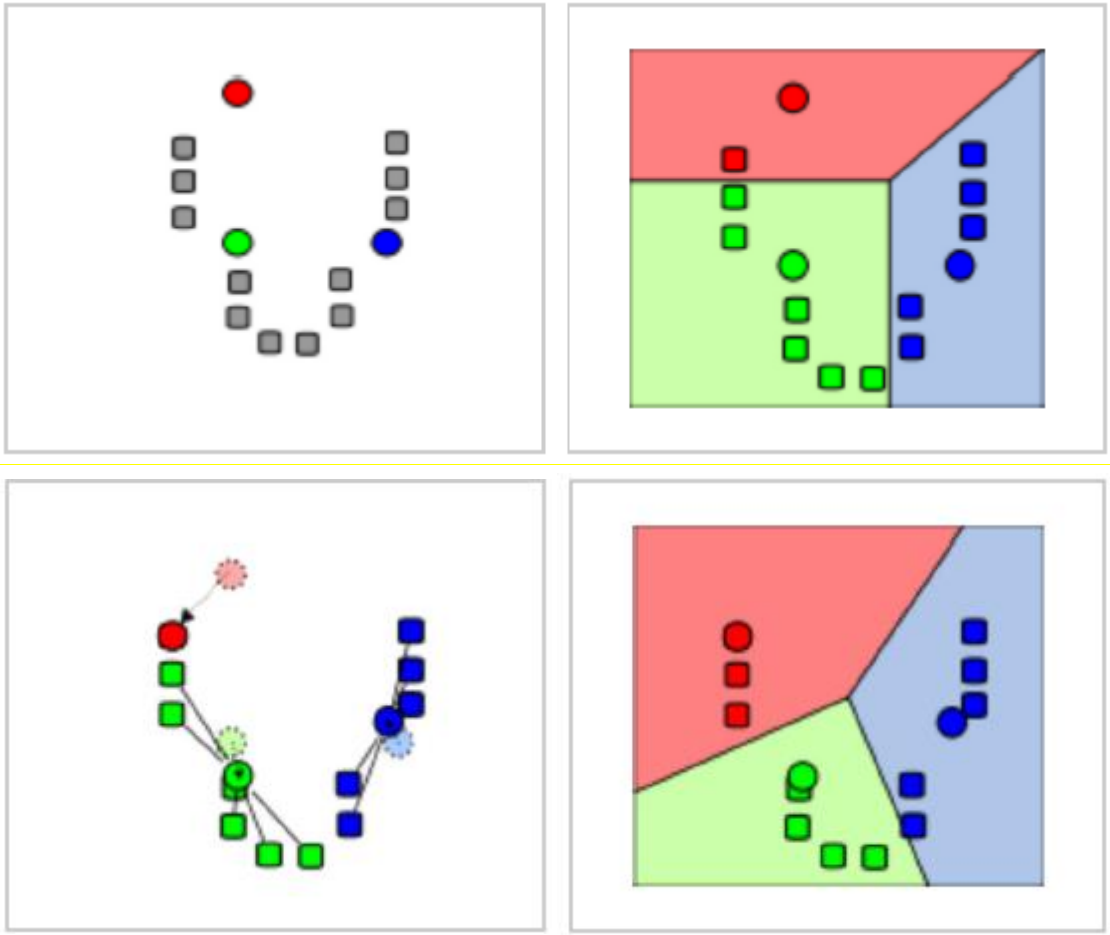
Алгоритъмът k-means (*k* средни) с входен параметър *k* разделя дадено множество от *n* обекта на *k* клъстера по такъв начин, че вътре-клъстерното сходство да бъде голямо, а между-клъстерното сходство – малко. Клъстерното сходство се измерва по отношение на средната стойност на обектите в клъстера – изкуствено създаден обект, наречен *центроид*, представляващ център на клъстера, който може да се разглежда като център на тежестта на този клъстер.

Алгоритмът започва работата си със случаен избор на *k* обекта като центрове на търсените клъстери. След това всеки от останалите обекти се разпределя към най-близкия клъстер, като разстоянието до клъстера се изчислява като разстояние от обекта до центъра на клъстера. След като всички обекти са разпределени, изчисляват се нови центрове на клъстерите, които представляват техни центроиди – обекти, за които стойността на всеки атрибут е средно аритметично от стойностите на този атрибут за обектите от съответния клъстер. След това описаният процес на разпределяне на обектите по клъстери се повтаря отново. Цикълът формиране на клъстери – уточняване на клъстерните центрове се изпълнява, докато бъде удовлетворен избраният критерий за качество на клъстеризацията.

Търсенето в *k*-means е ограничено до малка част от цялото пространство на възможните разделяния. По тази причина е възможно да бъдат изпуснати добри решения, ако алгоритъмът достигне локален екстремум за използвания критерий. За да се подобри качеството на решенията, често се прилага многократно повтаряне на алгоритъма с различни начални точки, случайно избрани в качеството на центрове на отделните клъстери. Окончателното решение е това, за което избраният критерий има най-добра стойност.

Алгоритъмът *k*-means може да бъде прилаган само към обекти, описвани с *непрекъснати атрибути*, тъй като само за такива е определено понятието средна стойност. Този алгоритъм работи добре при клъстери със сферична форма, относително еднакви по размер и добре разделени в пространството. Той не е подходящ за намиране на клъстери с произволна форма и много различаващи се по размер. Освен това, той е чувствителен към шумовете и крайностите, тъй като дори малък брой такива обекти може съществено да промени координатите на центроида.

Пример



Вариантът на *k*-means, позволяващ работа с номинални атрибути, се нарича *k-modes.* Основната идея на този метод е да бъде заменено Евклидовото разстояние с друга мярка за разстояние/различие, дефинирана върху номинални атрибути, и да бъде дефинирано понятието за средна стойност за такива атрибути.

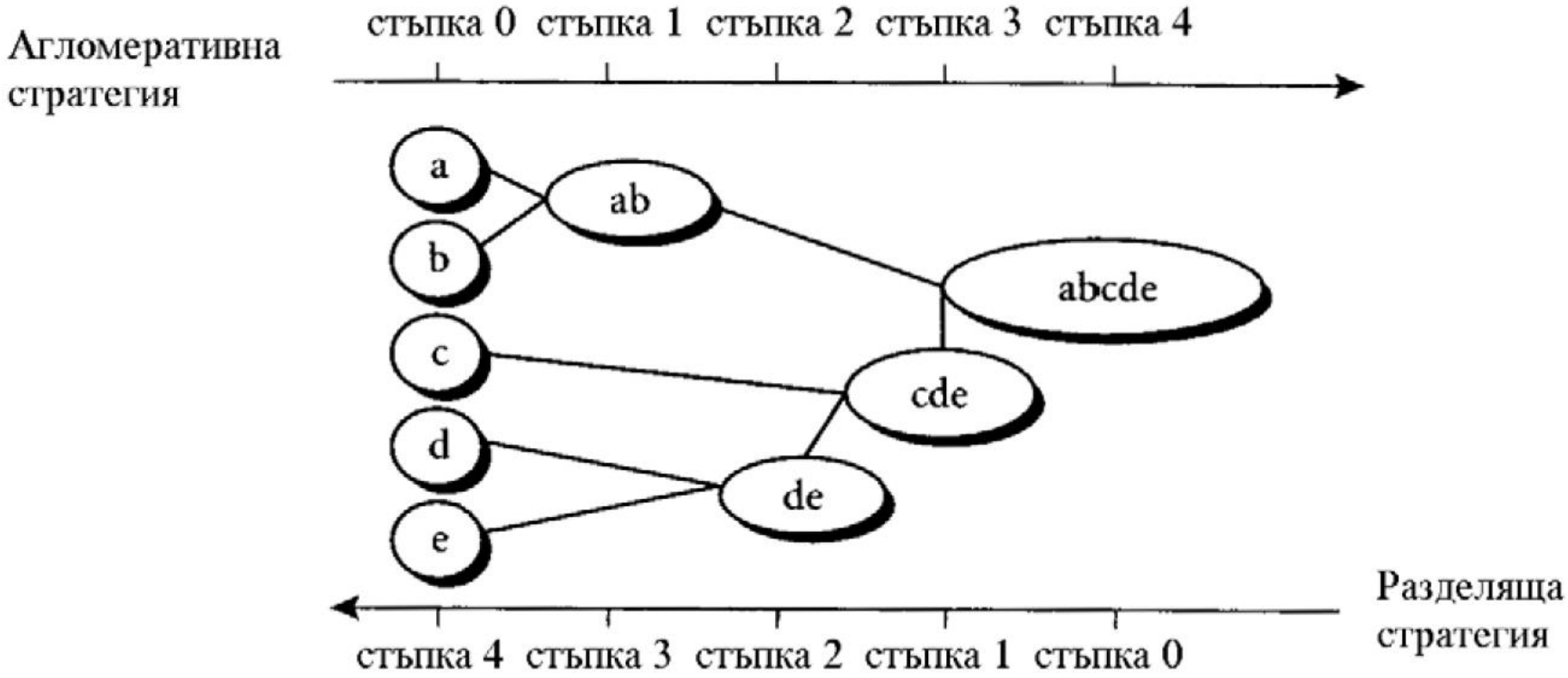
Методи за йерархична клъстеризация

Всеки метод за йерархична клъстеризация работи чрез групиране на дадените обекти в дърво от клъстери. Тези методи могат да бъдат разделени на два класа – *агломеративни* („слепващи“) и *разделящи* (от англ. divisive) в зависимост от това, дали йерархичната декомпозиция се извършва отдолу нагоре или отгоре надолу.

Агломеративна и разделяща йерархична клъстеризация

Методите за *агломеративна йерархична клъстеризация* започват работата си с поместване на всеки обект в отделен клъстер, след което така конструираните атомарни клъстери последователно се слепват във все по-големи и по-големи клъстери, докато всички обекти бъдат поместени в един единствен клъстер или бъде изпълнен предварително избраният критерий за прекратяване на работата на метода. Методите от тази група се различават по дефиницията на функцията на сходство между клъстери.

Методите за *разделяща йерархична клъстеризация* работят в посока, обратна на агломеративната клъстеризация – те започват работата си с поместване на всички обекти в един единствен клъстер. След това те итеративно „разбиват“ клъстера на все по-малки и по-малки части, докато бъдат получени атомарни клъстери (т.е. съдържащи само по един обект) или бъде изпълнен предварително избраният критерий за прекратяване на работата на метода, като например достигане на определен брой на клъстерите или достигане на определен праг за най-голямото разстояние между два(та) най-близки клъстера.



Йерархична клъстеризация на множеството обекти {a,b,c,d,e}

И двете стратегии за йерархична клъстеризация се базират в общия случай на избора на *метод за определяне на разстояние между клъстери*. Съществуват различни дефиниции на това разстояние, но всички те се базират на съответен начин за изчисляване на разстояния между двойки обекти от различни клъстери. Една от най-ранните и най-важните мерки за разстояние е *минималното разстояние*, което още се нарича *разстояние до най-близкия съсед* или *метод на единичното свързване (single link)*. Тя дефинира разстоянието между два клъстера като разстоянието между двойката най-близки обекти от тези клъстери.

Използването на тази мярка за разстояние между клъстери води до така наречения *„верижен ефект“,* при който дългите верижки от близо намиращи се точки (обекти) се причисляват към един и същ клъстер. Това означава, че описаният метод на единичната връзка има доста ограничено приложение за сегментиране. Освен това, той е чувствителен към малки промени в данните и към наличието на екстремни обекти. Методът на единичната връзка има едно свойство, уникално в сравнение с всички останали мерки за разстояние между клъстери: ако две двойки клъстери се намират на едно и също разстояние помежду си, то няма значение в какъв ред те ще се сливат (или разделят) – крайният резултат ще бъде един и същ.

Друга, отново екстремна мярка за разстояние между клъстери, е *максималното разстояние*, което още се нарича *разстояние до най-далечния съсед* или *метод на пълното свързване (complete link)*. То дефинира разстоянието между два клъстера като разстоянието между двойката най-отдалечени един от друг обекти от тези клъстери. Използването на тази мярка за разстояние води до създаване на клъстери с еднакъв размер в термините на обема на заеманото от тях пространство (не на броя на обектите в тях).

По средата между тези две екстремни мерки се намират *разстоянието между центроиди* и *средното разстояние.*

Основните проблеми при методите за йерархична клъстеризация са свързани с избора на точки за сливане или разделяне на клъстерите, тъй като след извършване на избраната операция всички следващи стъпки ще се извършват върху получените на текущата стъпка клъстери. Няма възможност както за връщане назад и промяна на вече взети решения, така и дори за размяна на обекти между вече създадени(те) клъстери. По този начин едно не най-добро решение за сливане или разделяне на клъстери, взето на някоя стъпка, може до доведе до ниско качество на клъстеризацията като цяло.