Клъстеризация на данни

- *Може ли човек да мисли?*
  Компютърен въпрос

- Клъстерен анализ (*Data clustering*) е задача за разделяне на дадена извадка от обекти (процеси) на непресичащи се еднородни множества, наречени клъстери, така че всеки клъстер да се състои от подобни обекти, а обектите от различните клъстери съществено да се различават. Задачата за клъстеризация се отнася към широк клас задачи за обучение без учител.

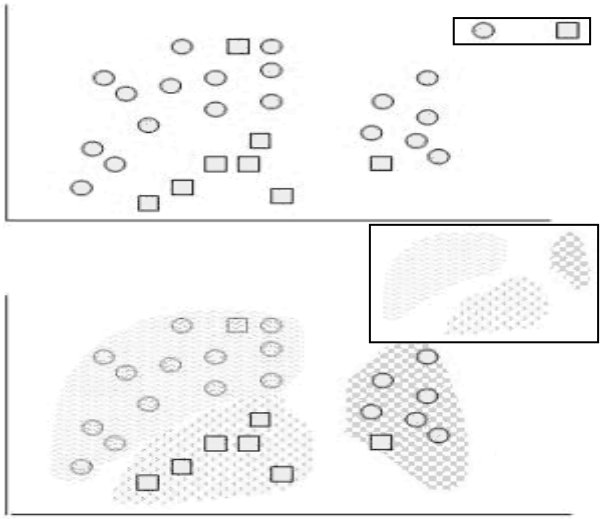
Същност

- Целта на клъстеризацията е търсене на съществуващи структури. Тя е сравнителен анализ. Описателна процедура е и не извършва никакви статистически изводи. Дава възможност да се проведе предварителен (разузнавателен) анализ и да се изучи "структурата на данните."
- Задачата за клъстеризация е сходна със задачата за класификация и е нейно логическо продължение. Разликата е в това, че класовете данни не са предварително определени. Синоним на термина "клъстеризация" е "автоматична класификация", "обучение без учител" и "таксономия" – йерархична класификация на обекти. Клъстеризацията (обучение без учител) се различава от класификацията (обучение с учител) по това, че изходното множество (клъстерите) обикновено не е известно.
- Терминът клъстерен анализ е въведен за първи път от Трион (Tryon) през 1939 год. За разлика от задачата за класификация, клъстерният анализ не изисква предварителни предположения за данните, не налага ограничения на представянето на изследваните обекти, позволява да се анализират показателите на различни типове данни. Клъстеризацията е групировка на обектите (наблюдения, събития) на основата на свойства (променливи), описващи същността на тези обекти. Колкото повече обектите вътре в клъстера са по-подобни един на друг и се отличават от обектите в другите клъстери, толкова по-точна е клъстеризацията. Променливите се измерват в сравними скали. Чрез клъстерния анализ може да се съкращава размерността на данните, като те се правят по нагледни.

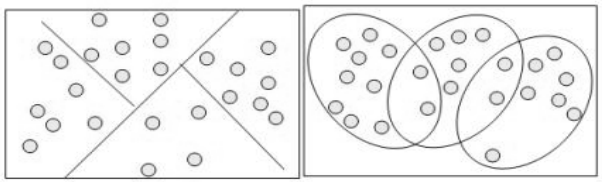
- Понятието "клъстер" (cluster) се превежда като струпване, грозд. Характеризира се като група от обекти, имащи общи свойства. Клъстерите се характеризират по два признака:
  - ✓ вътрешна еднородност;
  - ✓ външна изолираност.
- Полезен е, когато се класифицира голяма съвкупност от информация.

Сравнение на класификации и клъстеризация.

Характеристика	Класификация	Клъстеризация
Контролируемост на обучението.	Контролируемо обучение.	Неконтролируемо обучение.
Стратегия.	Обучение с учител.	Обучение без учител.
Наличие на маркировки на класа.	Обучаващо множество се съпровожда с маркировки, указващи класа, към който се отнася наблюдението.	Маркировките на класа на обучаващото множество са неизвестни.
Причина за класификация.	Да се класифицират нови данни на основата на обучаващо множество.	Установяване на съществуващи класове (клъстери) от дани.



Класификация и клъстеризация.



Непресичащи и пресичащи се клъстери.

Математическа постановка

$X = (x_1, x_2, ..., x_n)$ - множество от обекти

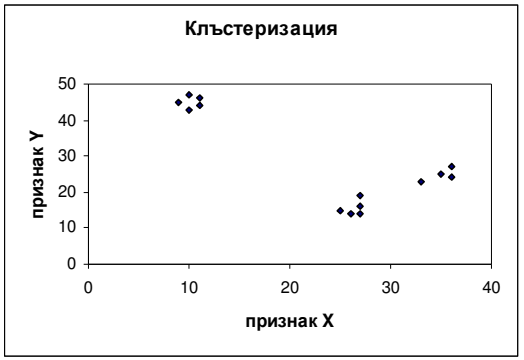
$Y = (y_1, y_2, ..., y_l)$ - множество от клъстери

$D(x_i, x_j)$ - функция на разстоянието между обектите

$X_m = (x_1, x_2, ..., x_m)$ - крайна извадка от обучаващи примери от множеството *X*

$x_i \in X_m$, y_j - номер на клъстер

$l \le n, m$ - брой клъстери

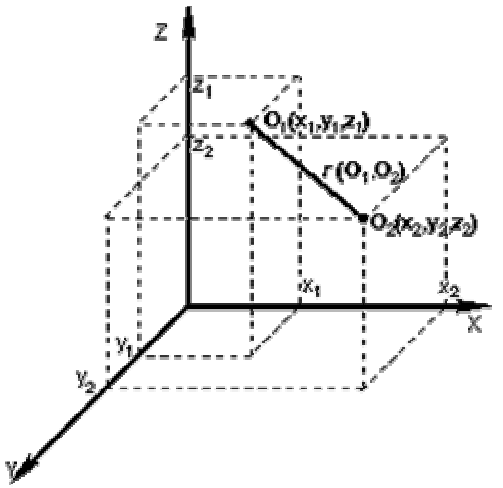


Набор от данни.

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Признак X	27	11	25	36	35	10	11	36	27	26	9	33	27	10
Признак Y	19	46	15	27	25	43	44	24	14	14	45	23	16	47

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

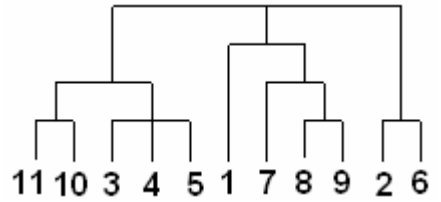
$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}$$



Разстояние между две точки в пространство с три измерения.

Методи на клъстерния анализ

- йерархически;
- нейерархически.
- *Йерархическата* клъстеризация последователно обединява малки клъстери в големи или разделя големи клъстери на по-малки. Те не изискват предварителни предположения за броя клъстери.



Пример на дендрограма.

нейерархически методи, основани на итеративни методи за разделяне на входната съвкупност от данни. В процеса на делението нови клъстери се формират, докато се изпълни правило за преустановяване на делението.
Съществуват два подхода:
✓ определяне границите на клъстерите, като най-плътните участъци в многомерното пространство от данни, тоест сгъстяване на точки;
✓ минимизация на мерките (размерите) на различните обекти.

Алгоритъм на k-средни (k-means)
Метод на най-близкия съсед или единична връзка
Метод на най-отдалечените съседи

Етапи на клъстерния анализ

Клъстерният анализ има следните етапи:

- Изключват се част от данните или се прави извадка от целият набор от данни;
- Избира се метрика за стандартизация на входните данни;
- Определят се количеството клъстери (за итеративния клъстерен анализ);
- Определя се метода за клъстеризация (правила за обединение, свързване) - решаващ е за определяне формата и характеристиките на клъстерите.

Невронни мрежи

Структурите, имащи свойствата на мозъка и нервната система, имат следните особености:

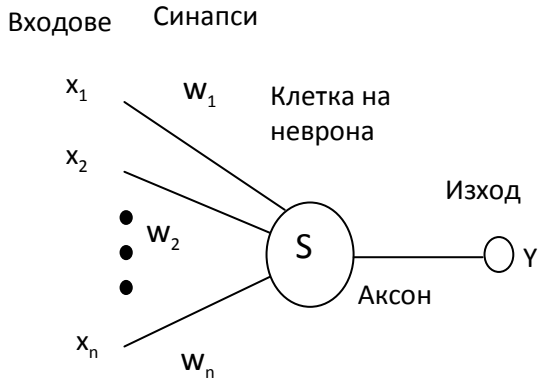
- Паралелна обработка на информацията;
- Способност за обучение;
- Способност за автоматична класификация;
- Висока надеждност;
- Асоциативност.

Изкуствен неврон

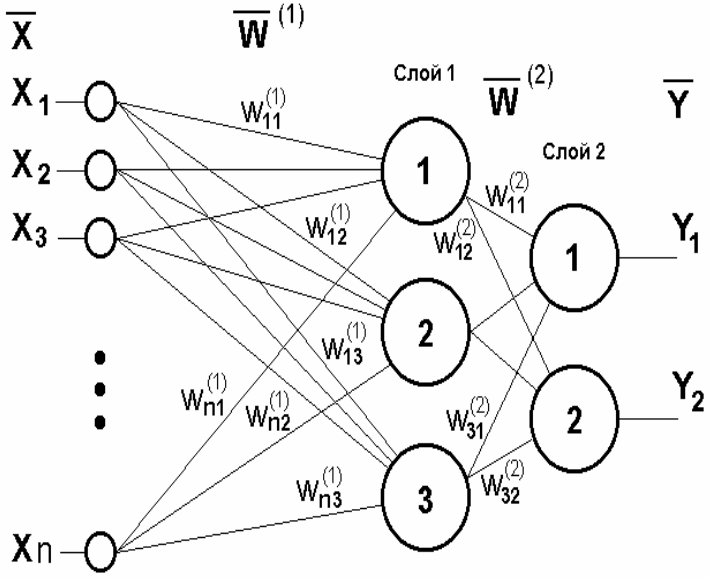
$$S = \sum_{i=1}^{n} x_i w_i$$

$Y = f(S)$ - активационна функция

$$f(x) = \frac{1}{1+e^{-ax}} \qquad f(x) = 1/\left(1+e^{-0x}\right) = 0{,}5$$



Основни елементи на неврона.



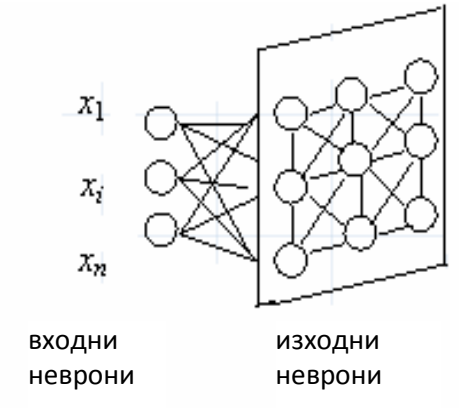
Двуслойна невронна мрежа.

Самоорганизиращи се карти на Кохонен

Самоорганизиращите се карти на Кохонен (Self Organized Map – SOM) са разновидност на невромрежовите алгоритми.

$$\vec{w} = [w_1, w_2, ..., w_n]$$

- Резултатът зависи само от структурата на входните данни.
- Два слоя – входен и изходен (слой на Кохонен).
- На невроните от входния слой се подават вектори на признаците на обектите за въвеждане в клъстерите. Броят на входните неврони е равен на размерността на векторите на атрибутите (броят атрибути на обекта).
- Броят на изходните неврони на мрежата на Кохонен е равен на броя клъстери, които трябва да създадат модела и всеки неврон се асоциира с определен клъстер. Изходният слой се обработва на принципа на „победителят взема всичко” - резултатът на неврона с най-голяма стойност на изхода е единица, а останалите са 0. Обектът се отнася към клъстера, асоцииран с единичния неврон (победителя).
- Невронните мрежи от подобен тип са подходящи за решаване на различни задачи във финансовата сфера.



Мрежа на Кохонен.

Алгоритъм на SOM

- *Инициализация на картата* – начално задаване на теглата на векторите за възлите.
- Определяне броя съседи на неврона-победител и обучение – изменение на векторите на теглата на неврона-победител и на неговите съседи с цел приближаване към наблюдението.
- Определяне с помощта на функция на съседствата *h* съседите на неврона-победител M_c . Функцията определя „мярката на съседство" на възлите M_c и M_i и изменение на вектора на теглата. Тя постепенно уточнява техните стойности, отначало за голямо количество възли по-бързо и след това за по-малко количество по-бавно. Като функция на съседствата обикновено се използва гаусовата функция на съседствата:

$$h_{ci} = \alpha(t) \cdot \exp\left(-\frac{\left| r_c - r_i \right|^2}{2\sigma^2(t)} \right)$$

$0 < \alpha(t) < 1$ -- обучаващ параметър, монотонно намаляващ с всяка следваща итерация (определя приближението на стойностите на вектора на теглата на неврона-победител и неговите съседи - по-голяма стъпка, по-малко уточнение);

r_i u r_c са координати на възлите $M_i(t)$ и $M_c(t)$ на картата;

$\sigma(t)$ - намалява количеството съседи с итерациите, монотонно намаляващ е.

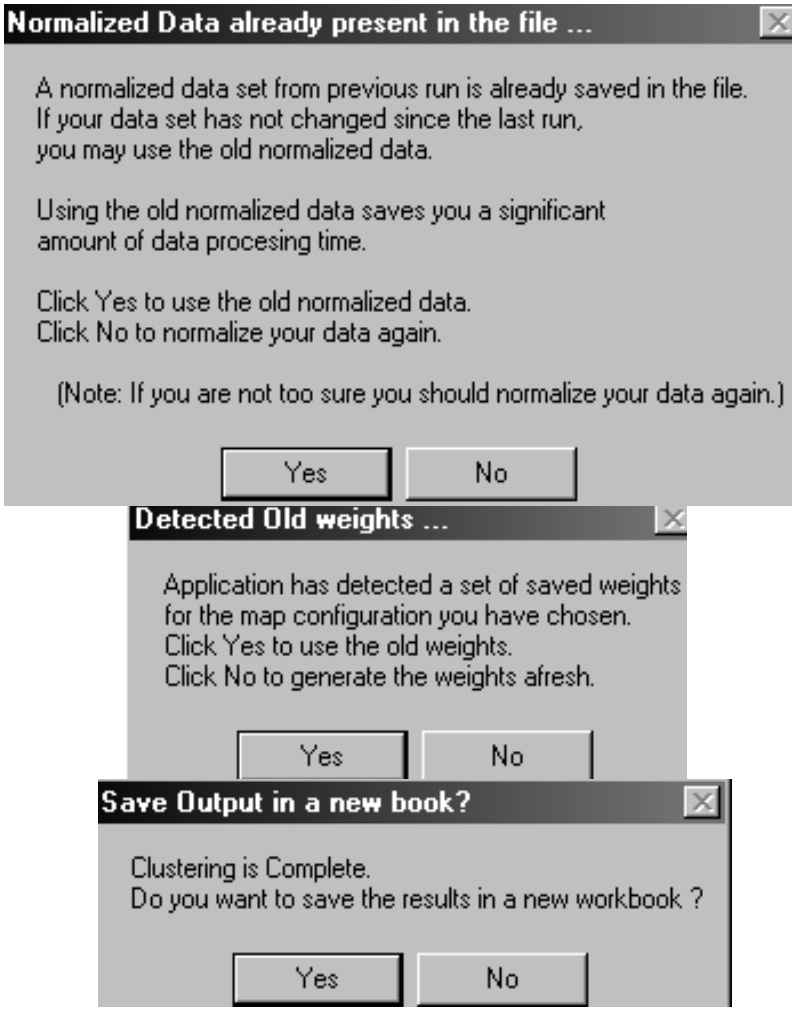
Параметрите α u σ и техният характер на намаляване се задава от аналитика.

Инструкции за ползване на NNclust.xls приложение

Стъпка 1: *Въвеждане на данни*

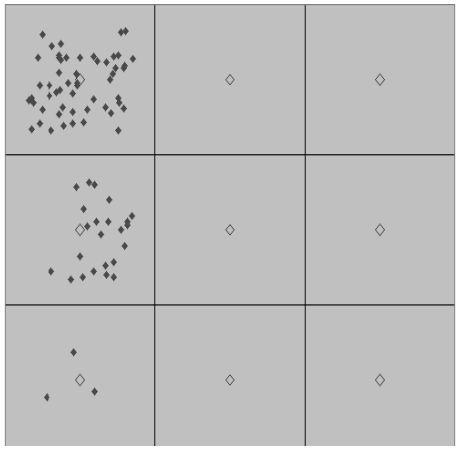
Стъпка 2: *Попълване на електронния лист* ***Input***

Стъпка 3: *Използване на бутон* ***Build Clusters***

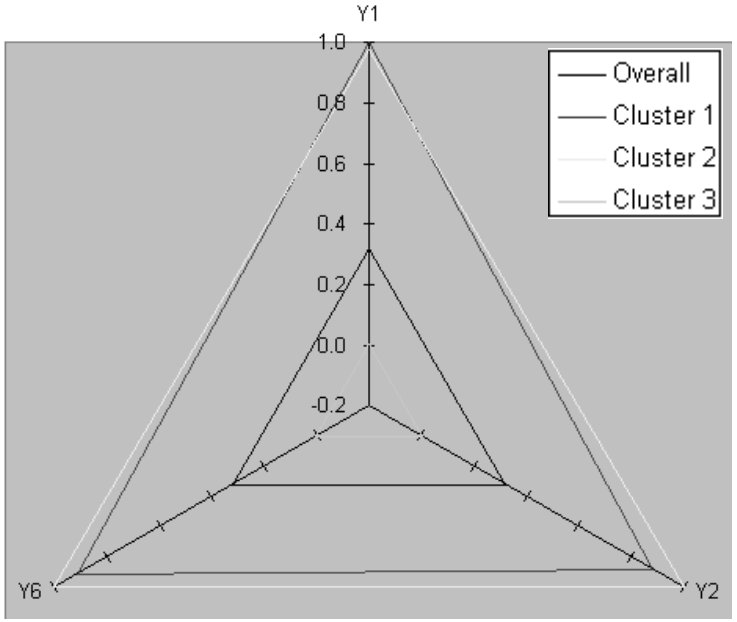


Етапи при създаване модела на клъстеризация.

Стъпка 4: *Резултати от клъстеризацията*



Диаграма на клъстерите.



Визуално сравнение средните стойности на клъстерите в *RadarPlot*.

Cluster Means

	Overall	Cluster 1	Cluster 2	Cluster 3
Y1	20,7	30,2	29,8	16,3
Y2	20,6	28,4	30,0	16,1
Y6	20,5	28,8	30,0	16,1

Средни аритметични стойности на входните променливи за клъстерите.

Cluster Variances

	Overall	Cluster 1	Cluster 2	Cluster 3
Y1	58.4	0.4	1.5	26.1
Y2	58.3	0.6	1.1	24.1
Y6	59.6	0.3	0.5	25.4

Стойности на дисперсията на входните променливи за клъстерите.

Положителни черти и недостатъци на приложението NNclust.xls

Положителни черти на приложението NNclust.xls,1:

- Колоните за вход и изход на променливите не е необходимо да са последователни. Например, може да има 20 колони за данните, а да се ползват 2,3,5 и 10 за клъстеризация. Не е нужно да са съседни;
- Може да се сменят някои от обучаващите параметри и да се изследва влиянието им върху резултатите от клъстеризацията;
- Докато се извършва обучението на картата, данните се въвеждат в *случаен ред*;
- След като картата се обучи, приложението записва векторите на теглата на картата. Следващият път, когато ще се обучава картата със същатата размерност и същите данни, ще има опция за използване на теглата, които вече са записани като стартиращи тегла. Това подобрява обучението на картата;
- Приложенитео има някои възможности да управлява липсващи данни. Всяка нечислова стойност в произволна колона на данните, които се използват в клъстеризирането, ще се третира като липсваща стойност. Приложението ще замести всички липсващи стойносит в колоната със средната стойност;

Положителни черти на приложението NNclust.xls,2:

- След като клъстеризацията приключи, на изхода ще се изведат средната стойност и дисперсията на всеки клъстер. Отбелязва се разположението му върху картата. Приложението също генерира картата, заедно с точките на данните върху нея, за да изобрази разположението на клъстерите;
- След като обучението приключи, моделът може да се запише в отделна работна книга. Данните, заедно с етикетите на клъстерите също се записват. Записват се и променливите, които не се ползват за клъстеризация. Това може да е полезно за профилиране на клъстерите. Например, за клъстеризация на потребителски данни се ползват променливи – доход, кредит, месечни разходи. Обаче базата от данни може също да има променлива възраст. След завършване на клъстеризацията, би трябвало да се провери разпределението по възраст на всеки клъстер. Например дали някои клъстери съдържат преимуществено млади хора;
- Ако се избере опцията за запис на модела в отделна работна книга, може също и да се изисква диаграмата *RadarPlot*. Тази диаграма визуално сравнява средните стойности на различните клъстери.

Някои важни недостатъци

- Ограничения на размера – приложението може да управлява най-много 50 входни променливи. Приложението не може да управлява категориални данни. Всички колони за използване в клъстеризацията трябва да са числови. Клетки, които са празни, текстови или с грешки, трябва да се махнат. В противен случай ще се търсят всички клетки с нечислови данни и ще се заместват със средната стойност. Това отнема много време;

- Могат да се стартират максимум 500 обучаващи цикъла. Това не е голямо ограничение, понеже може да се стартира от теглата, до които е достигнато и да се стартира за нови 500 цикъла и т.н.;

- Скоростта е важна при създаване на клъстерите. За умерено големи съвкупности от данни приложението работи бавно. Скоростта се определя основно от:
  - ✓ Предварителната обработка на данните преди обу-че-ние-то да стартира;
  - ✓ Крайното изчертаване на картата след като обу-че-ние-то е завършило.

Някои тънкости

- Приложението използва квадратна мрежа от неврони, разположена в редове и колони. То не може да управлява неквадратна мрежа;
- Приложението ползва гаусова функция на съседствата;
- Теглата на векторите на картата се инициализират с координатни стойности между -1 и 1;
- Входните променливи са мащабирани да приемат стойности между -1 и 1;
- Използвана е Евклидовата дефиниция на разстоянието. Приложението не може да управлява други видове разстояния.