Системи, основани на знания - зимен семестър, 2020/2021 учебна година

Тема 15: Клъстеризация

Какво представлява клъстеризацията?

Клъстеризация е дейността по групиране на набор от обекти (вектори или точки), така че обектите от една и съща група (наричана клъстер) да са подобни помежду си и да се различават от обектите в останалите клъстери.

На пръв поглед терминологията на клъстера изглежда ясна и точна: *група от сходни обекти*.

Но клъстерите, открити от различни алгоритми, могат да се различават значително по своите свойства. Следователно, съществен е изборът на подходящ алгоритъм при разрешаването на конкретен проблем.

Моделите за клъстеризация се разделят и на „hard” и „soft”. При „hard” клъстеризацията всеки обект попада в точно един клъстер. При „soft” клъстеризацията всеки обект може да попадне в повече от един клъстер, като за всеки клъстер, в който попада, обектът притежава съответна степен на принадлежност.

Алгоритъм *k*-means

Сравнително прост алгоритъм. Стреми се да раздели обектите на *k* (*k* е дадено естествено число) клъстера по следния начин:

- По случаен начин се избират центровете на всички клъстери;
- Повтарят се следните стъпки:
  - всеки обект се асоциира с (причислява към) клъстера с най-близък център;
  - замества се всеки център на клъстер със средното на всички обекти, асоциирани с него.

Описание на алгоритъма

Приемаме, че данните са в двумерното пространство и образуват k клъстера.

По случаен (напълно произволен) начин избираме k обекта $m_1^{(1)}$, … ,$m_k^{(1)}$, които наричаме *средни*.

Разпределяне

Добавяме всяка точка (всеки обект) x_p към клъстера с/около най-близкото средно. Така получаваме текущото множество от клъстери:

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \ \forall j, 1 \leq j \leq k\}$$

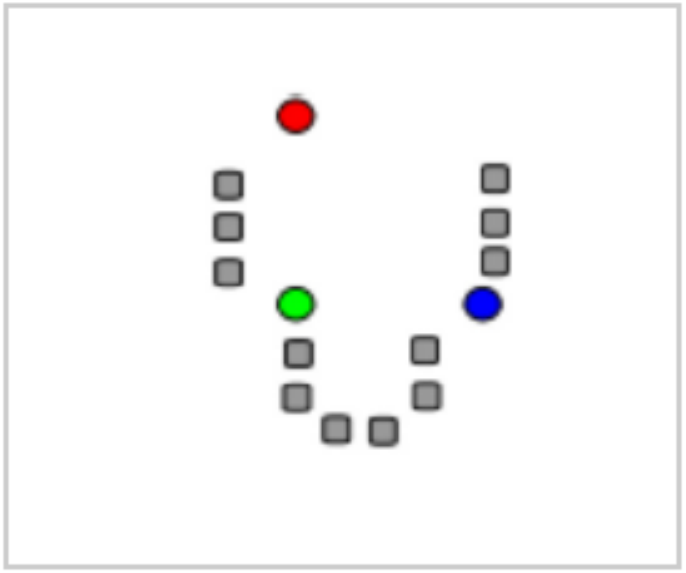
При това всяка точка x_p се асоциира с точно един клъстер $S^{(t)}$, дори и ако може да бъде асоциирана с два или повече клъстера.

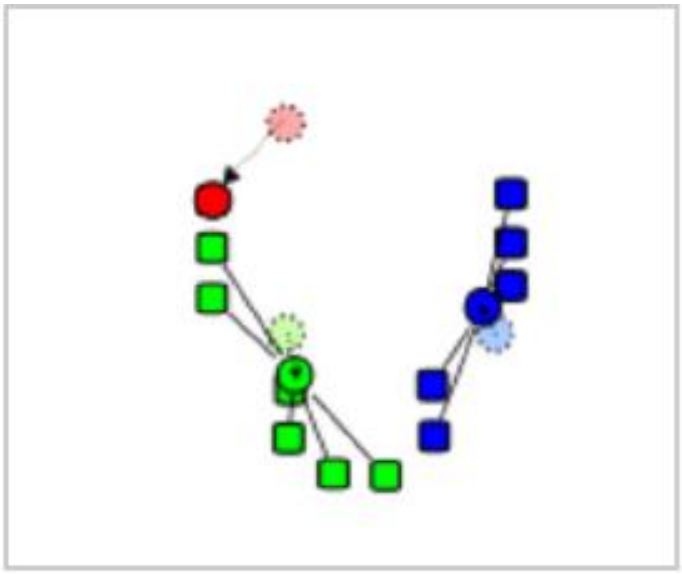
Обновяване

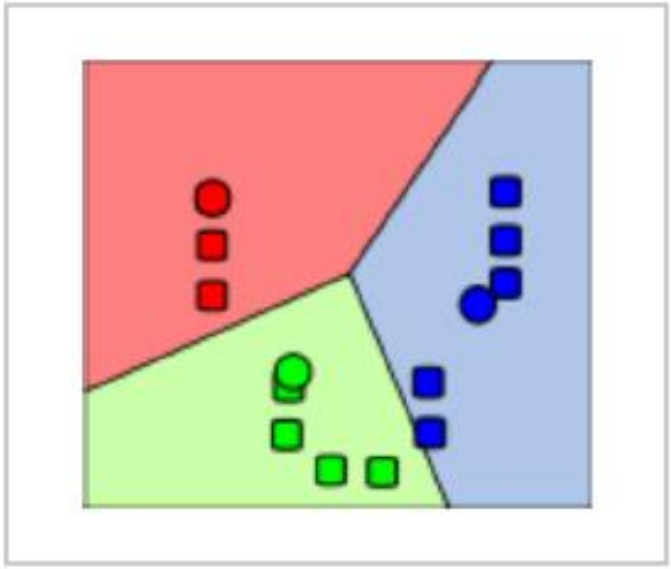
Обновяваме множеството от средните, като за нови средни избираме центроидите на новопостроените клъстери:

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

Пример







Оценка на алгоритъма

Алгоритмът *k*-means представлява техника за локално търсене с цел оптимизиране на разпръскването на данните.

Тъй като това е евристичен алгоритъм, няма гаранция, че ще се получи оптимално групиране. Резултатът може да зависи от избора на началните средни.

Алгоритъмът обикновено е много бърз, но има случаи, при които дори в двумерно пространство, може да отнеме експоненциално време $2^{\Omega(n)}$.

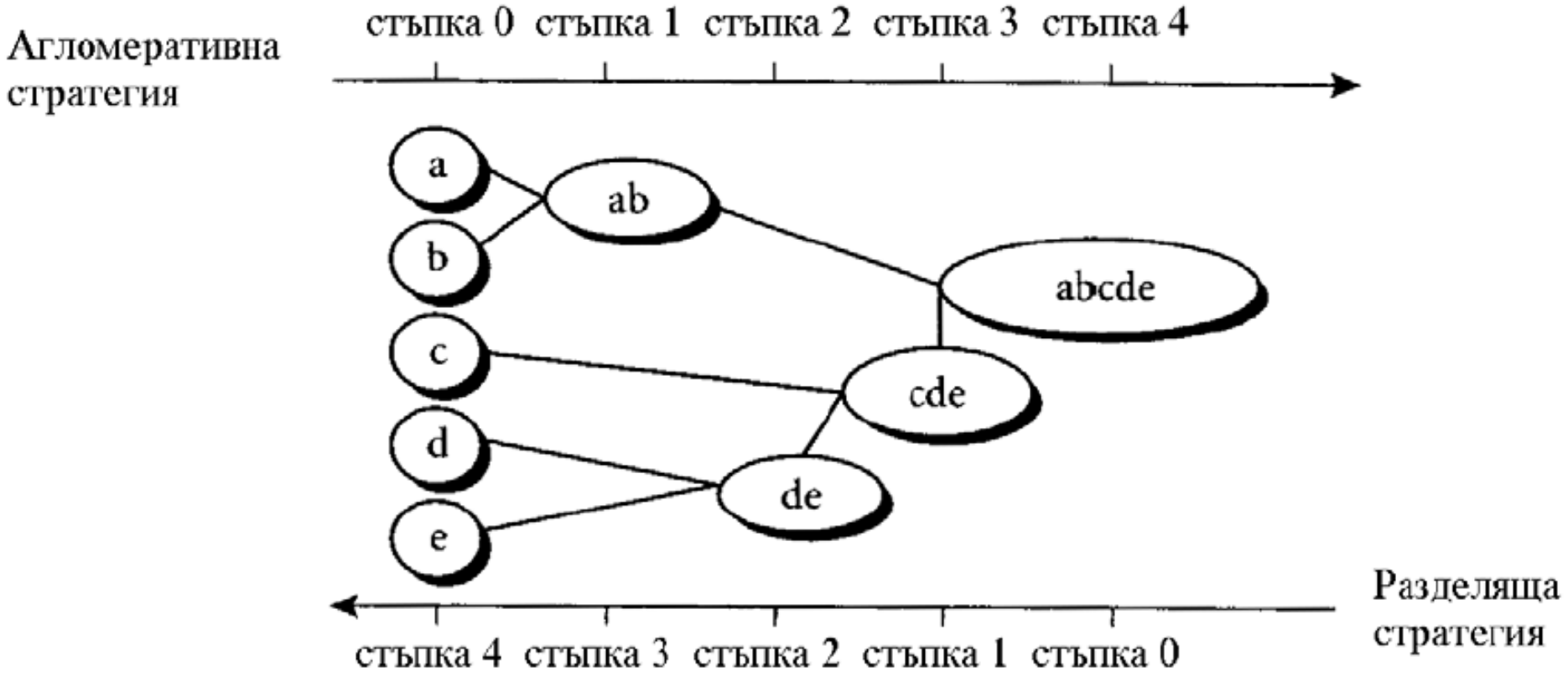
Такива случаи обаче рядко се срещат в практиката.

Йерархично клъстериране

Йерархичното клъстериране е метод, който създава йерархия от клъстери. За разлика от *k*-means, не е необходимо да зададем броя клъстери (*k*). Необходимо е само да посочим по какъв начин ще се сравняват обектите (например използваното разстояние). Като резултат получаваме дърво, листата на което са построените клъстери.

Има два типа йерархично клъстериране:

- *агломеративен* – подхожда „отдолу нагоре“, всеки обект стартира като собствен клъстер и с изкачването нагоре клъстерите се групират един с друг;
- *разделящ* – подхожда „отгоре надолу“, всички обекти стартират като един клъстер и със слизането надолу рекурсивно се разделят.


Агломеративна стратегия
стъпка 0 стъпка 1 стъпка 2 стъпка 3 стъпка 4
a
b
c
d
e
ab
de
cde
abcde
Разделяща стратегия
стъпка 4 стъпка 3 стъпка 2 стъпка 1 стъпка 0

Резултатът от прилагането и на двете стратегии за йерархично клъстериране се базира на избрания метод за определяне на *разстоянието между клъстери*.
Съществуват различни дефиниции на това разстояние, като всички те характеризират с определен начин за изчисляване на разстоянията между двойки обекти от различни клъстери.

Една от най-ранните и най-популярните мерки за разстояние между клъстери е *минималното разстояние*, което още се нарича *разстояние до най-близкия съсед.* Тя дефинира разстоянието между двата клъстера като разстояние между двойка(та) най-близки обекти от тези клъстери.

В такъв случай се говори за т. нар. *метод на единичното свързване* (*single-link clustering*).

Използването на тази мярка за разстояние между клъстери води до така наречения *„верижен ефект“*, при който дългите верижки от близо намиращи се точки (обекти) попадат в един и същ клъстер.

Методът на единичното свързване има едно важно свойство: ако две двойки клъстери се намират на едно и също разстояние помежду си, то няма никакво значение в какъв ред те ще се сливат или разделят – резултатът ще бъде един и същ.