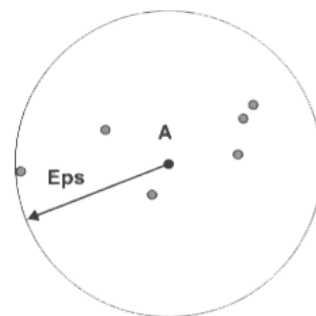


Лекция 7. Методи за клъстеризация базирани на плътността

Методите, базирани на плътността (гъстота) (*density-based*) са разработени за намирането на клъстери с произволна форма. Те разглеждат клъстери като области в пространство с голяма плътност, т.е. с голяма гъстота на обекти, които са разделени чрез области от пространството с по-малка плътност, които се третират като шум. Съществуват множество клъстерни алгоритми, базирани на различни интерпретации на понятието „плътност”. Ще започнем с алгоритъма DBSCAN, който разглежда плътността като брой на точки в околността на всяка точка със зададен радиус. След това ще разгледаме подход към клъстеризацията, базиран на използване на мрежи. При този подход цялото пространство се разбива на клетки от мрежата, а клъстерите се формират от съседни клетки с достатъчна плътност. След това ще разгледаме алгоритъм DENCLUE – една техника за клъстеризация, която моделира плътността като сума от влияния на така наречени функции за разпределение на плътност, свързани с всяка отделна точка. На края ще се запознаем с алтернативното понятие за мярката на сходство, която се базира на броя на общи най-близки съседи на всяка точка, както и с дефиницията на плътността, базираща се на този нов поглед на близост между точките. Ще бъдат представени и два алгоритъма за клъстеризация, използващи тези понятия.

7.1. DBSCAN: Метод за клъстъризация, базиран на свързаните области с достатъчно голяма плътност

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) е алгоритъм, който може да намира клъстери с произволна форма в бази от пространствени, зашумени данни, като ги построява от области с достатъчно голяма плътност на обекти (точки) (M. Ester et al. 1996). Алгоритъмът използва дефиницията за плътността, която се основава върху понятието за централна точка. По-точно, плътността се оценява за всяка отделна точка от зададеното множество точки чрез изчисляване на броя на точки, лежащи в околността с радиус Eps около тази точка (така наречена *Eps-околността*). В тази бройка се включва и самата точка. Този подход към определяне на плътността е илюстриран на Фиг. 7-1. Броят на точки в околността на точка А с радиус Eps е равен на 7, включвайки и самата точка А.



Фиг. 7-1. Понятието за плътност

7.1.1. Основни понятия

Съгласно формулираната по-горе дефиниция на плътността всяка точка от множеството, подлежащо на клъстеризация, може да бъде класифицирана като:

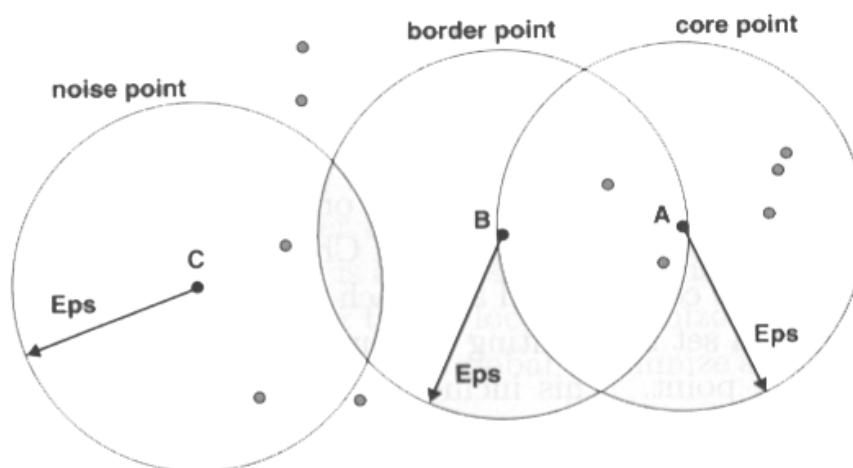
- а) Принадлежаща към вътрешността на някоя област с голяма плътност (*вътрешна (core) точка*)
- б) Принадлежаща към границата на някоя област с голяма плътност (*гранична (border) точка*)
- в) Принадлежаща към област с малка плътност (*шум*)

По формално:

Вътрешни точки: това са точки, намиращи се във вътрешността на един базиран на плътността клъстер. Една точка е вътрешна, ако броят на точки в нейната *Eps*-околност превишава определен, зададен от потребителя праг *MinPts*. Например, на Фиг. 7-2 точка *A* е вътрешната точка за указания радиус *Eps*, ако $MinPts \leq 7$.

Гранични точки: Една гранична точка не е вътрешна точка, но принадлежи към околността на някоя вътрешна точка. На фиг. 7-2 точка *B* е граничната точка. Една гранична точка може да попада в околността на няколко вътрешни точки.

Шум: Една точка се класифицира като шум, ако тя е нито вътрешна, нито гранична точка. На Фиг. 7-2 точка *C* е шум.



Фиг. 7-2. Вътрешна точки, гранична точка и шум.

7.1.2. Алгоритъм DBSCAN

Неформално алгоритмът DBSCAN може да се опише по следния начин. Всяка двойка вътрешни точки, които са достатъчно близки, т.е. в *Eps*-околности една на друга, се поместват в един и същ клъстер. Всяка гранична точка, която е

достатъчно близко до някоя вътрешна точка, се помещава в същия клъстер като и вътрешната точка (Ако една гранична точка е разположена на едно и също разстояние до вътрешните точки от различни клъстери, прилага се процедура по решаване на конфликти). Шумът се премахва. По-формално описание на алгоритъма е приведено в таблица 7-1. Алгоритъмът използва същите понятия и намира същите клъстери както и оригиналният DBSCAN, но е направен с оглед на по-голяма простота от колкото ефективност.

-
1. Маркирай всички точки като вътрешни, гранични или шум.
 2. Изтрий шума.
 3. Свържи с дъги всички вътрешни точки, намиращи се в предели на разстоянието Eps една от друга (таква точки още се наричат *плътностно-свързани*).
 4. Превърни всяка група от плътностно-свързани вътрешни точки в отделен клъстер.
 5. Назначи всяка гранична точка към един от клъстери асоциирани със съответните вътрешни точки
-

Таблица 7-1. Алгоритъм DBSCAN

При използване на пространствения индекс, изчислителната сложност на DBSCAN е $O(n \log n)$, където n е броят на обекти в базата данни. В противен случай тя е $O(n^2)$. Алгоритъмът прави нейерархичната клъстеризация и е чувствителен към параметри, задавани от потребителя.

7.1.3. Избор на параметри

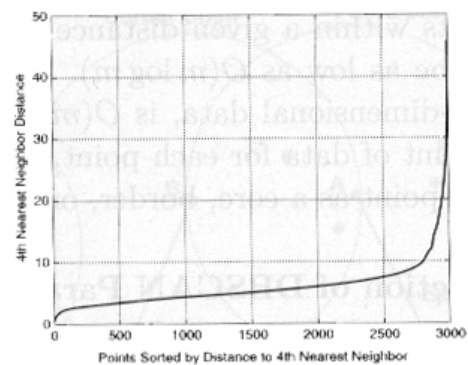
Базовият подход към избора на параметрите на алгоритъма се състои в анализа на поведение на разстояние от всяка точка до нейния k -ти най-близък съсед – така наречено k -разстояние. Предполага се, че за точки, принадлежащи към един и същ клъстер, стойността на k -разстояние ще бъде малка, ако k не е по-голямо от размера на клъстера. Естествено е, че възможни определени вариации в зависимостта от плътността на клъстера и случайното разпределение на точки, но в средно, диапазонът на такива вариации няма да е голям, ако плътностите на различни клъстери не са радикално различни. Обаче за точки, намиращи се извън клъстера, т.е. шум, стойността на k -разстояние ще бъде относително голяма. По тази причина, ако изчислим k -разстояния за всички точки при избраната стойност на k , да ги сортираме в нарастващ ред, а след това да ги нанесем на графика, то на него можем да очакваме да видим една рязка промяна в стойността на k -разстояние, която отговаря на една подходяща стойност на параметъра Eps . Ако изберем тази стойност като Eps и вземем стойността на k като стойността на параметъра $MinPts$, то точки, за които k -разстоянието е по-малко от Eps , ще бъдат маркирани като вътрешни точки, докато останали – като гранични точки или шум.

На Фиг. 7-3 е показано едно примерно множество от точки, за които е построен график на k -разстояния, показан на Фиг. 7-4. Стойността на Eps , определяна по такъв начин, зависи от стойността на k , но не се променя значително при промяната на k . Ако стойността на k е прекалено малка, то дори един малък брой на близко разположени в пространство точки, които са шум или крайности, ще бъдат некоректно маркирани като клъстери. Ако стойността на k е прекалено голяма, то малки клъстери (с размер по-малък от k) са вероятно да бъдат маркирани като шум. Оригиналният алгоритъм DBSCAN използва стойността $k = 4$, което изглежда една приемлива стойност на повечето от двумерни множества данни.



Фиг. 7-3. Примерни данни

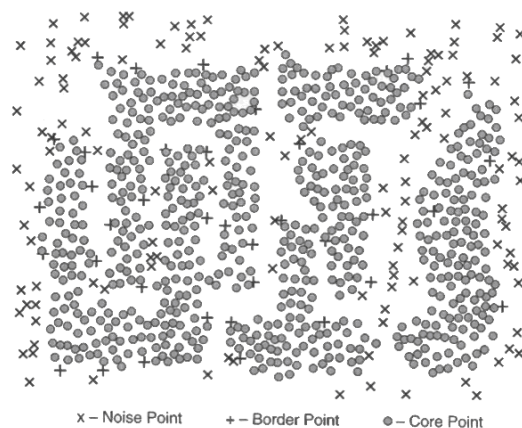
Пример 7.1 На Фиг. 7-3 е показано сравнително сложно множество от 3000 двумерни точки. Параметър Eps за тези данни е бил определен чрез графика на сортираното разстояние до 4-ти най-близък съсед за всяка точка (Фиг. 7-4). С помощта на графика е била определена стойност, където се наблюдава рязкото нарастване на измерваната величина. Била е избрана стойност $Eps = 10$, отговаряща на точка на прелома на кривата. Клъстерите, намерени от DBSCAN при тези параметри, т.е. $MinPts = 4$ и $Eps = 10$ са показани на Фиг. 7-5 (a). Вътрешните точки, граничните точки и шум са показани на Фиг. 7-5 (b).



Фиг. 7-4. График на k -разстояние за примерните данни



(a) Клъстери, намерени от DBSCAN



(b) Вътрешни точки, гранични точки и шум

Фиг. 7-5. DBSCAN клъстеризация на 3000 двумерни точки

7.2. Клъстеризация, базирана на мрежи

Мрежата е един ефективен начин за организиране е едно множество от данни, особено с малка размерност. Идеята е да бъдат разбити стойности на всеки атрибут на определен брой непрекъснати интервали, които да образуват множество от клетките на мрежата (става дума за атрибути с наредба). Всеки обект попада в тази клетка, в която съответния атрибутен интервал съдържа стойностите на обекта. Обектите могат да бъдат разпределени по клетки само с едно минаване чрез всички данни, като по същото време за всяка клетка може да бъде събрана информацията за броя на точки в клетката.

Има няколко метода за клъстеризация с използване на мрежи, но повечето от подходите се базират на плътността. В този раздел ще използваме термин „клъстеризация, базирана на мрежи” да означим клъстеризацията, базирана на плътността, използваща мрежи.

Базовият алгоритъм за такава клъстеризация е показан в Таблица 7.2.

-
1. Определи множество от клетки на мрежата.
 2. Назначи обекти към съответните клетки и изчисли плътността на всяка клетка.
 3. Премахни клетки, в които плътността е под зададения праг τ .
 4. Формирай клъстери от непрекъснати (съседни) групи от плътни клетки.
-

Таблица 7-2. Базовия алгоритъм за клъстеризация, базирана на мрежа

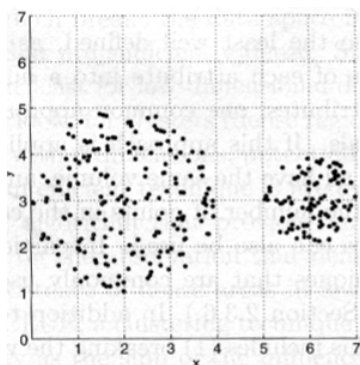
7.2.1. Определяне на клетки на мрежата и тяхната плътност

Тази ключова стъпка на алгоритъма е ясно определена, тъй като съществуват множество различни алгоритми за разбиване на множеството възможни стойности на всеки атрибут на определен брой непрекъснати интервали. Например, за непрекъснати атрибути най-често се използва дискретизацията на интервали с еднаква ширина. Ако този подход се прилага към всеки атрибут, то получените клетки на мрежата ще имат еднакъв обем и, следователно, плътността на една клетка е просто броят на обекти в нея.

Освен този подход могат да бъдат приложени и вече ни известните подходи за дискретизация на интервали с еднаква дълбочина или с използване на клъстеризация. Плътността на една клетка се дефинира като броят на точки в клетката, делен върху нейния обем, не зависимо от размерността на използваното пространство.

Пример 7.2. На Фиг. 7-6 са показани два множества от двумерни точки, разбити на 49 клетки чрез използване на 7 по 7 мрежата. Първото множество съдържа 200 точки, генерирани от еднородното разпределение в окръжността с център (2, 3) и радиус 2, докато второто множество е генерирано от еднородното разпределение в окръжността с център (6, 3) и радиус 1. Броят на точки за всяка клетка от мрежата е

указан в Таблица 7-3. Тъй като клетките имат еднакъв обем (площ), можем да разглеждаме тези стойности като стойностите на плътности на клетките.



Фиг. 7-6. Плътността, базирана на мрежа

0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

Таблица 7-3. Брой точки в клетките на мрежата

7.2.2. Формиране на клъстери от плътни клетки на мрежа

Формирането на клъстери от съседни групи от плътни клетки е относително праволинейно (например, от Фиг. 7-6 е ясно, че има само два клъстера). Обаче, има няколко въпроса за решаване. Първо, трябва да определим, какво разбираме под съседни клетки. Например, дали една клетка на двумерната мрежа има 4 съседни клетки или 8? Второ, ние се нуждаем от ефективни техники за намиране на съседните клетки, особено когато само попълнени (т.е. съдържащи обекти) клетки се помнят.

Алгоритъмът за клъстеризация, описан в Таблица 7.3, има няколко ограничения, които могат да бъдат преодоляни чрез малко усложняване на самия алгоритъм. Например, възможно е да има частично празни клетки по границите на клъстери. Често тези клетки не са плътни и, следователно, те ще бъдат премахнати, като по този начин частите от клъстера ще бъдат загубени. Фиг. 7-6 и Таблица 7-2 показват, че четири части от по-големия клъстер ще бъдат изгубени, ако прагът за плътността бъде установен на 9. Процесът на клъстеризация трябва да бъде модифициран, за да избегне премахването на такива клетки, макар че това ще изисква допълнителни изчисления.

Базовият алгоритъм за клъстеризация с използване на мрежа може да бъде подобрен ако освен плътността, може да се ползва и друга, допълнителна информация. В много случаи данните имат както пространствени, така и непространствени атрибути. С други думи, някои от атрибутите описват разположение на обекти в пространство или време, докато други - описват други аспекти на обектите. Един често срещан пример е къщи, които имат както местоположение, така и много други характеристики от рода на цена, жилищната площ и т.н. Поради пространствена (или времева) автокорелация, обектите в една отделна клетка често имат сходни стойности на тези атрибути. В такива случаи е възможно клетките да бъдат филтрирани на базата на някои статистически характеристики на един или няколко непространствени атрибути, например

средната цена на къща, а след това да бъдат формирани клъстери на базата на плътността на останалите точки.

Положителната черта на клъстеризация, базирана на мрежи, е нейната ефективност. При зададено разбиване на всеки атрибут на интервали, само чрез едно обхождане на данните е възможно за всеки обект да бъде определена клетката, както и за всяка клетка – броя на обекти в нея. Освен това, макар броят на потенциални клетки на една мрежа може да бъде голям, клетките трябва да бъдат създавани само за непразни клетки. По този начин изчислителната и пространствената сложност на процедурата по определяне на мрежата, назначаване на обекти в клетки и изчисляване на плътността на всяка клетка е само $O(n)$, където n е броят на точките. Чрез използване на специални методи за индексирание (от типа на дървото на търсене) съседните плътни клетки могат да бъдат ефективно намирани, което прави процесът на клъстеризация много ефективен, например $O(n \log n)$. По тази причина описания по-горе базовият алгоритъм за клъстеризация чрез плътността с използване на мрежи е послужил като основа за голям брой ефективни клъстерни алгоритми, такива като STRING, GRIDCLUS, WaveCluster, CLIQUE и т.н.

Отрицателната черта на клъстеризацията, базирана на мрежи, както и на повечето алгоритми, базирани на плътността, се състои в голямата зависимост от избора на прага за плътност τ . Ако прагът е много висок, част от клъстери могат да се загубят. Ако той е прекалено нисък, някои от двойки клъстери, които трябва да бъдат разделени, ще останат слети. Освен това, ако има клъстери и шум с различна плътност, то е възможна ситуация, при която не е възможно да бъде намерена една единствена стойност на прага τ , която работи добре във всички части на пространството от данни.

Освен това съществуват ред допълнителни проблеми с клъстеризацията, базирана на мрежи. Например (виж Фиг. 7-6) правоъгълната мрежа не хваща точно плътността на области с кръгли граници. Можем да опитаме да избегнем този проблем чрез създаване на една по-фина мрежа, но тогава броят на точки, асоциирани с една клетка, могат да имат големи колебания, тъй като точките не са разпределени равномерно. Дори някои от клетките, дори разположените вътре в клъстера, могат да останат празни. Освен това, в зависимост от разположение и размер на клетките, една група точки може да се окаже само в една единствена клетка или да бъде разбита на няколко различни клетки. В първия случай тази група може да формира част от клъстера, докато във втория – може да бъде премахната.

И на края, при нарастване на размерността, броят на потенциални клетки на мрежата бързо нараства – експоненциално относително броя на размерностите. Макар че не е необходимо явно да разглеждаме празни клетки, е напълно възможно да се случи, че повечето от клетките на мрежата ще съдържат само по един обект. С други думи клъстеризацията, базирана на мрежи, има тенденция да работи лошо при данни с голяма размерност.

7.3. DENCLUE: клъстеризация, базирана върху функции за разпределяне на плътността

DENCLUE (DENSity-based CLUstering) е метод за клъстеризация, базиран върху множество от функции за разпределяне на плътността (A. Hinneburg et al. 1998). Методът се базира върху следните идеи:

1. Влиянието на всяка точка (обект) може да бъде формално моделирано чрез използване на определена математическа функция, наречена *функция на влияние*, която описва въздействието на точката вътре в своята околност.
2. Общата плътност на пространството от точки може да бъде моделирана аналитично като сумата от функциите на влияние на всички точки.
3. Клъстерите могат да бъдат определени математически чрез идентифициране на *точки за привличане на плътността (density attractors)*, които са точки на локален максимум на общата функция на плътността.

Нека x и y са обекти в d -мерното пространство на признаци (атрибути) F^d . Функцията на влияние, което оказва обект y върху x е функция $f_B^y : F^d \rightarrow \mathbb{R}^+$, отобразяваща пространството на признаци в пространството на реални положителни числа, и се дефинира в термини на базовата функция на влияние:

$$f_B^y(x) = f_B(x, y) \quad (7.1)$$

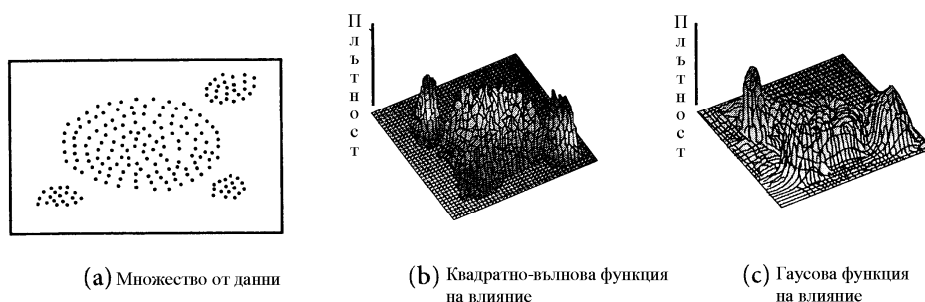
В общия случай функцията на влияние може да бъде произволна функция, определена като функцията от разстояние между два обекта в някоя околност. Използвайки за функцията на разстояние Евклидовото разстояние $d(x, y)$, може да се дефинира следните две базови функции на влияние:

$$\text{Квадратно-вълнова функция на влияние } f_{\text{Square}}(x, y) = \begin{cases} 0, & \text{ако } d(x, y) > \sigma \\ 1, & \text{в противен случай} \end{cases} \quad (7.2)$$

$$\text{Гаусова функция на влияние } f_{\text{Gauss}}(x, y) = e^{-\frac{d(x, y)^2}{2\sigma^2}} \quad (7.3)$$

Функцията на плътност на някой обект $x \in F^d$ се дефинира като сумата от функциите на влияние на всички обекти от множеството на данни D :

$$f_B^D(x) = \sum_{x_i \in D} f_B^{x_i}(x) \quad (7.4)$$



Фиг. 7-7. Възможни функции на плътност за едно двумерно множество от точки

Изображение на разпределението на функцията на плътността за едно множество от двумерни точки е показано на фиг. 7-7.

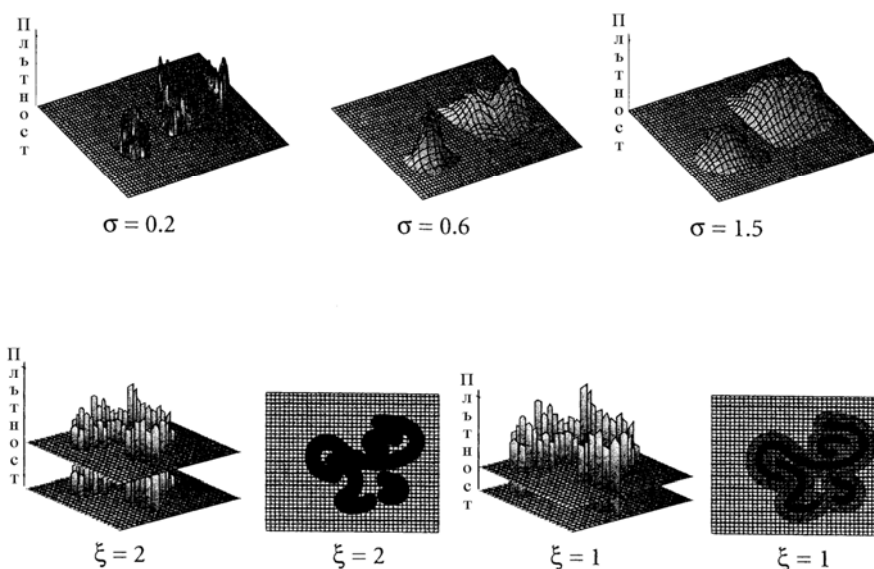
От функцията на плътността могат да бъдат определени *градиентът* на функцията и нейните локални максимуми – *точки на привличане на плътността*. Ще казваме, че точка x се *привлича по плътността* към точката за привличане на плътността x^* , ако съществува множество от точки $x_0, x_1, \dots, x_k$ такива, че $x_0 = x$, $x_k = x^*$ и градиентът на x_{i-1} е в направление x_i за $0 < i < k$. За една непрекъсната и диференцируема функция на влияние точката за привличане на плътността на едно множество от данни може лесно да бъде намерена чрез използване на метода на градиентното спускане (най-бързото спускане).

Използвайки тези понятия могат формално да бъдат дефинирани клъстери, определени чрез своя център, и клъстери с произволна форма:

За дадената точка на привличане на плътността x^* , един *клъстер, дефиниран чрез център*, е подмножеството от точки C , които са привличани към x^* и функцията на плътността на x^* е не по-малка от зададения праг ξ (т.е. $f_B^D(x^*) \geq \xi$); в противен случай (т.е. когато $f_B^D(x^*) < \xi$) клъстерът се разглежда като екстремната стойност (крайност).

Един *клъстер с произволна форма* (т.е. дефиниран чрез няколко центъра) е множество C_s от дефинирани чрез центрове клъстери, които са свързани чрез някой път от точки, чиято функция на плътността е не по-малка от прага ξ .

Примери на дефинирани чрез център клъстери и клъстери с произволна форма са показани на фиг. 7-8.



Фиг. 7-8. Примери на централно дефинирани клъстери (горния ред) и клъстери с произволна форма

DENCLUE има ред предимства в сравнение с други алгоритми за клъстеризация:

1. Алгоритъмът има солидна теоретична база и представлява обобщение на такива методи за клъстеризация, като:
 - a. *Клъстеризация, базирана на плътността*. Например, DBSCAN може да се моделира чрез прилагане на квадратно-вълнова базова функция на влияние и формиране на клъстери с произволна форма. В този случа параметър σ играе ролята на ε , а $\xi = \text{MinPts}$.
 - b. *Клъстеризация чрез разделяне*. Например, алгоритъм k-Means се моделира с Гаусова базова функция на влияние и използване на клъстери, дефинирани чрез център. В този случай $\xi = 0$, а σ се определя по такъв начин, че да бъдат формирани k клъстера.
 - c. *Йерархична клъстеризация*. Дефинираните чрез център клъстери за различни стойности на σ формират определена клъстерна йерархия.
2. Методът има добри клъстерни свойства за множество данни с голямо ниво на шума. Това се обяснява с факта, че точките на притегляне на плътността не се променят с нарастване на нивото на шума.
3. Методът позволява едно компактно математическо описание на произволни по форма клъстери, съставени от точки с голяма размерност
4. Методът работи значително по-бързо от някои други известни алгоритми (например той е около 45 пъти по-бърз от DBSCAN), тъй като работи върху пространствена решетка, организирана във вид на определена дървовидна структура.

Обаче, DENCLUE изисква много внимателен избор на параметъра σ и нивото на шума ξ , тъй като техните стойности могат значително да повлияят на качеството на резултати от клъстеризация.

7.4. Сходство, базирано на общи най-близки съседи

В някои случаи клъстерните техники, базирани се на стандартни подходи към определяне на сходството и плътност, не произвеждат желани клъстерни резултати. В този раздел ще проанализираме някои причини за това и ще въведем един непряк подход към оценката на сходството, който се базира на следния принцип:

Ако две точки са сходни с едно и също множество от други точки, то те са сходни една с друга, дори директната оценка на тяхното сходство не го показва.

7.4.1. Проблеми с традиционното сходство за данни с голяма размерност

В пространства с голяма размерност стойността на сходството между обекти обикновено е ниска. Да разгледаме като пример едно множество от документи – колекцията от статии, събрани от различни раздели на вестника: Развлечения, Финанси, Свят, Времето, Вътрешни новини и Спорт. Вие вече знаете, че тези

документи могат да бъдат представени като вектори в пространство с голяма размерност, където стойността на всеки атрибут на вектора указва броя на срещане на съответната дума от речника в дадения документ. За оценката на сходство между документи често се използва метриката на сходство, представляваща косинус на ъгъла между съответните вектори. В таблица 7-4 са показани средните стойности на тази мярка за всяка от описаните раздели на вестника Los Angeles Times, както и за цялото множество от документи.

Раздел	Средна стойността на косинус-сходство
Развлечения	0.032
Финанси	0.030
Свят	0.030
Времето	0.021
Вътрешни новини	0.027
Спорт	0.036
Всички раздели	0.014

Таблица 7-4. Сходство между документи в различни раздели на един вестник

Сходството на всеки документ със своя най-близък съсед е по-добро – в средно 0.39. Обаче, низката стойност на сходство между обекти от един и същ клас води до това, че често най-близкия съсед на един обект не принадлежи към същия клас. Например, в колекцията от документи, които са били използвани за създаване на Таблица 7-4, около 20% от документите имат своя най-близък съсед от друг клас (раздел). В общия случай, ако директното сходство е ниско, то се превръща в едно ненадеждно средство, определящо клъстеризация на обекти, особено в случая на агломеративна йерархична клъстеризация, където най-близките точки се сливат и след това никога не могат да бъдат разделени. Не зависимо от това, в повечето случаи едно голямо множество от най-близки съседи на един обект принадлежат към същия клас; този факт може да бъде използван за дефиниране на едно мярка на близост, която е по-подходяща за целите на клъстеризацията.

7.4.2. Проблеми с разлики в плътността

Другият проблем е свързан с разлики в плътността на различни клъстери. На Фиг. 7-9 е показана една двойка клъстери от двумерни точки с различна плътност. По-малката плътност на десния клъстер се отразява в по-голямо средно разстояние между точките. Дори когато точките в клъстера с по-малката плътност формират валиден клъстер, обикновените клъстерни техники срещат големи затруднения с намиране на подобни клъстери. Освен това, нормалните мерки за оценяване на компактността на клъстери, такива като, например, средно-квадратичната грешка, ще показват, че такива клъстери имат по-малка компактност. Да илюстрираме това твърдение с реален пример: звездите в една галактика са не по-малко реални клъстери от звездните обекти отколкото планетите в една слънчева система, макар че планетите в слънчевата система са значително по-близки една до друга в средно, отколкото звездите в една галактика.



Фиг. 7-9. Два кръгови клъстера от по 200 равномерно разпределени точки

7.4.3. Изчисляване на сходство, базирано на общи най-близки съседни

И в двата разгледани по-горе случая, решението е да отчитаме контекста, в който се намират точки, при изчисляване на сходството между тях. Тази идея може да бъде представена количествено чрез използване на определение за *сходство, базирано на общи най-близки съседни (SNN-сходство)*, което се изчислява по начина, указан в Таблица 7-5:

-
1. Намери k -най-близки съседни за всички точки.
 2. **Ако** две точки x и y не са сред k -най-близки съседни един на друг, **то**
 3. $similarity(x, y) \leftarrow 0$
 4. **Иначе**
 5. $similarity(x, y) \leftarrow$ брой на общи най-близки съседни
 6. **Край ако**
-

Таблица 7-5. Алгоритъм за изчисляване на SNN-сходство

На практика, SNN-сходството е броят на общи най-близки съседни, когато и двете точки са присъстват в списъка на най-близки съседни една на друга. Графично, това понятие е илюстрирано на Фиг. 7-10.



Фиг. 7-10. Изчисляване на SNN-сходство между две точки

Всяка от двете черни точки има по осем най-близки съседни, включително и самите точки. Четири от тези най-близки съседни – точки, показани със сиво – са общи. Следователно, SNN-сходството между тези две точки е равно на 4.

Графът на сходството между обекти, в който сходството се измерва като SNN-сходство, се нарича *графът на SNN-сходство*. Тъй като множество двойки от обекти ще имат стойността на SNN-сходството равно на 0, то един такъв граф ще бъде силно разпръснат (разреден).

7.4.4. SNN-сходство срещу директното сходство

SNN-сходството е полезно, тъй като решава проблеми, възникващи при използване на директното сходство. Първо, тъй като то отчита контекста на обекта чрез използване на броя на общи най-близки съседни, SNN-сходството дава решение в ситуации, когато един обект се намира относително близо до друг обект, принадлежащ към друг клас. В такива случаи обикновено обектите нямат много общи най-близки съседни и тяхното SNN-сходство е ниско.

SNN-сходството също така решава проблема с клъстери с различна плътност. В една област с ниска плътност обектите са по-далеч един от друг в сравнение с по-плътни области. Обаче, SNN-сходството на една двойка точки зависи само от броя на най-близките съседни, които са общи и за двете точки, а не от това, колко са далече тези съседни от всяка от точките. По този начин SNN-сходството извършва едно автоматично скалиране по отношение на плътността на тези точки.

7.4.5. Алгоритъм за клъстеризация на Jarvis-Patrick

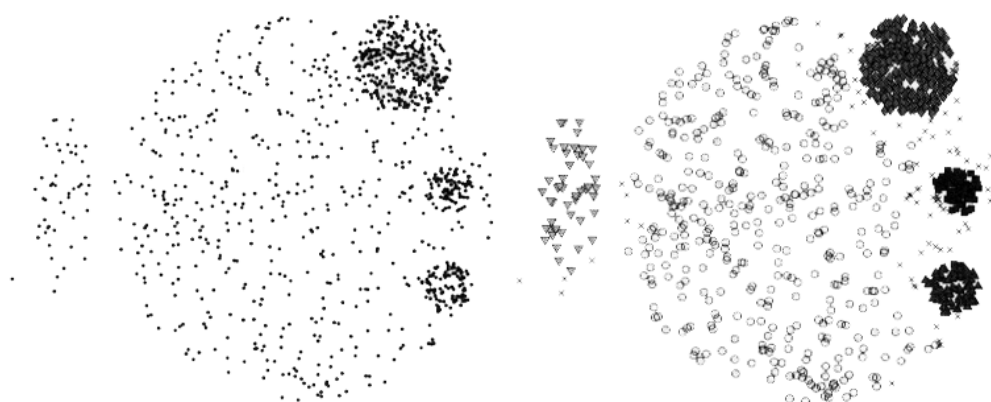
Алгоритъмът за клъстеризация на Jarvis-Patrick (JP-алгоритъм) илюстрира използването на понятието SNN-сходство (виж Таблица 7-6). Той заменя директното сходство между две точки с SNN-сходството, което се изчислява съгласно алгоритъма, показан в Таблица 7.5. След това за разреждане на матрицата на SNN-сходства между точките се използва един праг, чиято стойност се задава от потребителя. В термините на графи това означава, че се създава и след това се разрежда един граф на SNN-сходства. Клъстерите са просто взаимосвързани компоненти на този граф.

-
1. Изчисли графа на SNN-сходство.
 2. Разрежи полученият SNN-граф чрез прилагане на избрания праг за сходство.
 3. Намери взаимосвързани компоненти (клъстери) в разрежения SNN-граф.
-

Таблица 7-6. JP-алгоритъм за клъстеризация

Базовата изчислителна сложност на алгоритъма е $O(n^2)$, тъй като създаването на списъци с k-най-близки съседи изисква изчисляване на $O(n^2)$ разстояния (n е броя на точки).

Пример 7.3. Да разгледаме прилагане на JP-алгоритъма за клъстеризация на множеството от двумерни точки, показани на Фиг. 7-11 (a). Намерените клъстери са показани на Фиг. 7-11 (b). Размерът на списъка на най-близките съседи е бил 20, а двете точки се поместваха в един и същ клъстер, ако те са имали най-малко 10 общи най-близки съседи. Различните клъстери са показани с различни маркери и защриховане. Точките, маркирани със знака „x”, са били класифицирани от алгоритъма като шум. Те основно се намират в областите на прехода между клъстери с различна плътност.



(a) Оригинални данни (b) Клъстери, намерени от JP-алгоритъм
Фиг. 7-11. JP-клъстеризация на едно множество от двумерни точки

Тъй като JP-алгоритъмът се базира на SNN-сходство, той е добър за работа с шума и крайности, а също така е способен да намира клъстери с различни размери, форма и плътност. Алгоритъмът работи добре с данни с голяма размерност и особено добър в намиране на стегнати клъстери от силно свързани обекти.

Обаче JP-алгоритъмът определя един клъстер като един свързан компонент в SNN-граф, така че дали някое множество от обекти ще бъде разделено на два клъстера или оставено като един клъстер може да зависи само от една единствена връзка. Следователно JP-алгоритъмът за клъстеризация е прекалено чувствителен, т.е. той може да раздели истинските клъстери или да слее клъстери, които трябва да бъдат оставени разделени.

Едно друго потенциално ограничение на алгоритъма е че не всички обекти се клъстеризират. Обаче, тези обекти могат да бъдат добавени към съществуващите клъстери, а, в някои случаи, и не съществува изискването за клъстеризацията на абсолютно всички обекти. И на края, както в много други алгоритми за клъстеризация, изборът на най-добрите стойности на параметрите е доста сложна задача.

7.3. SNN-плътност

Както вече беше отбелязано по-рано, традиционната плътност, базирана на Евклидовото разстояние, става безсмислена при големи размерности. Естествено може да използваме дефиницията за плътност, базирана на централен обект (както в DBSCAN) с прилагане на някоя метриката за сходство, която работи добре с данни с голяма размерност – например косинус или Жакардов коефициент. Но, както вече беше отбелязано в точка 7.4.4, и тези мерки имат сериозни проблеми. Обаче, тъй като SNN-сходството отчита локалната конфигурация на точки в пространството на данни, то може да се използва като един обещаващ кандидат за дефиниране на една нова мярка за плътност.

В този раздел ще дефинираме понятието за SNN-плътност, следвайки подход, предложен в DBSCAN. Основните дефиниции в този раздел ще повтарят дефинициите от раздела 7.1 с подходящо модифициране, отчитайки, че ще използваме SNN-сходството.

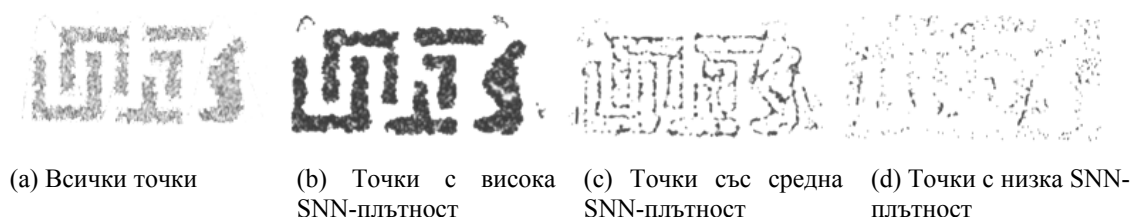
Вътрешни точки: Една точка е вътрешна, ако броят на точки в нейната *Eps*-околност, определена съгласно SNN-сходството и стойността на параметъра *Eps*, превишава определен, зададен от потребителя праг *MinPts*.

Гранични точки: Една гранична точка не е вътрешна точка, т.е. няма достатъчно точки в нейната *Eps*-околност, за да стане тя вътрешна точка, но самата точка попада в *Eps*-околността на някоя вътрешна точка.

Шум: Една точка се класифицира като шум, ако тя е нито вътрешна, нито гранична точка.

SNN-плътността измерва степента, в която една точка е окръжена от подобни точки (по отношение на своите най-близки съседи). Следователно, точките от области с висока и ниска плътност обикновено ще имат относително висока SNN-плътност, докато точките в области на прехода от ниска към голяма плътност, т.е. точки, лежащи между клъстери, ще имат ниска SNN-плътност. Подобният подход може да се окаже по-добре приспособен за данни, в които присъстват големи вариации в плътността, но клъстерите с малка плътност остават предмет на интереса.

Пример 7.4. На Фиг. 7-12 (а) са показани 10 000 двумерни точки. Фигури 7-12 (b-d) показват различието между тези точки, базирано на стойността на тяхната SNN – плътност – (b) са точки с най-висока SNN-плътност, (c) – със средна стойност на тази величина и (d) – с най-ниската стойност на SNN-плътност. От тези фигури може да се види, че точките с висока SNN-плътност (т.е. с най-висока свързаност на SNN-граф) са кандидати да бъдат вътрешни точки, тъй като те имат тенденция да са разположени във вътрешността на клъстера, докато точките с най-ниската SNN-плътност са кандидати за шум и крайности, тъй като повечето от тях са разположени в области, окръжаващи клъстерите.



Фиг. 7-12. SNN-плътност на двумерни точки

7.3.1. Клъстеризация базирана на SNN-плътност

За създаване на новия клъстеризационен алгоритъм дефинираната по-горе SNN-плътност може да се комбинира с алгоритъма DBSCAN. Този алгоритъм е подобен на JP-алгоритъм по това, че започва работата си с изчисляване на SNN-граф. Обаче, вместо използване на някаква прагова стойност на SNN-сходството за разреждане на графа и анализ на останалите свързани неговите компоненти, новият алгоритъм просто повтаря стъпките на DBSCAN (виж Таблица 7-7).

-
1. Изчисли граф на SNN-сходство
 2. Приложи DBSCAN алгоритъм с указани от потребителя параметри *Eps* и *MinPts*.
-

Таблица 7-7. Алгоритъм за клъстеризация, базиран на SNN-плътност

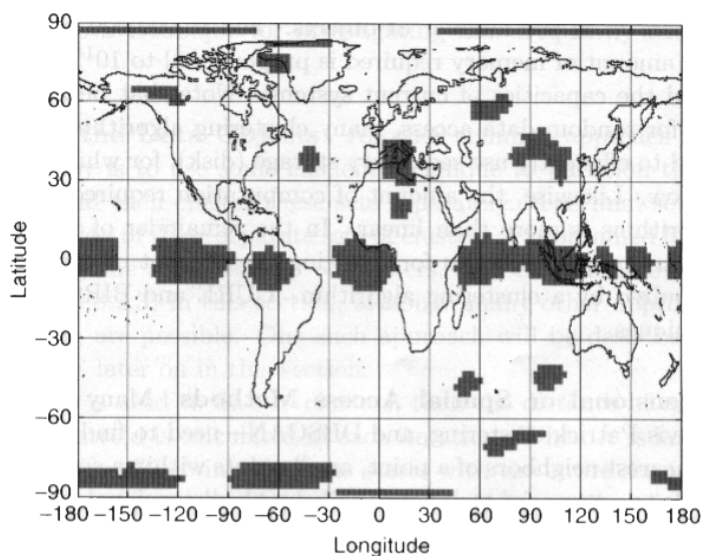
Алгоритмът автоматично определя броя на клъстери в данни. Не всички данни са клъстеризират – точките, които се игнорират, се разглеждат като шум или крайности, както и точки, които нямат силна връзка с някоя група от точки. Клъстеризацията, базирана на SNN-плътност, намира клъстери, в които точките са силно свързани една с друга. В зависимост от конкретното приложение алгоритмът може да игнорира голяма част от точки. Например, клъстеризацията, базирана на SNN-плътност, е много добра за намиране на теми в групи от документи.

Представеният алгоритъм е по-гъвкав от JP-алгоритъма за клъстеризация или от DBSCAN. В отлчието от DBSCAN той може да бъде използван при данни с висока размерност, както и в ситуации, когато клъстерите имат различна плътност. В отлчието от JP-алгоритъма, който използва прага, а след това определя останалите свързани компоненти на графа като клъстери, клъстеризацията, базирана на SNN-плътност, използва по-устойчив подход, основан на понятията за SNN-плътност и вътрешни точки.

Пример 7.5. (Базираната на SNN-плътност клъстеризация на времеви серии).

За илюстрация на възможностите на подхода ще покажем, как описаният алгоритъм може да бъде използван за клъстеризация на времеви серии, описващи месечни данни за атмосферното налягане в различни точки на Земята. По-точно, данните описват средно-месечното налягане на морското ниво (SLP) за период от

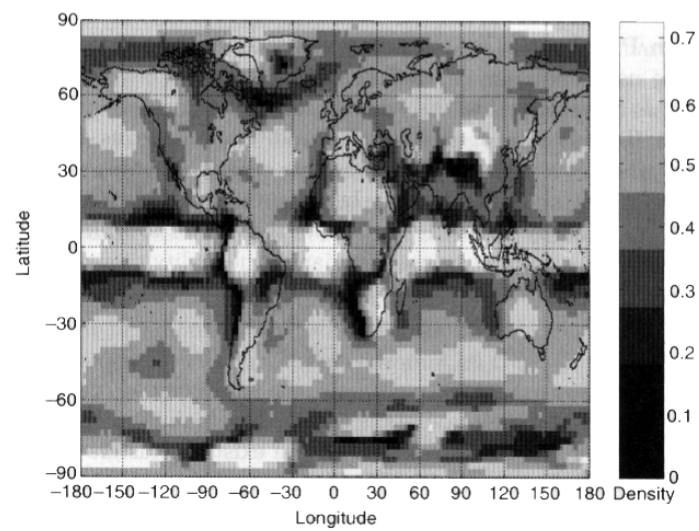
41 години във всяка точка от мрежа – дължина-широчина - със стъпка 2.5° . Клъстерният алгоритъм намери клъстери (сивите области), показани на Фиг. 7-13. Обърнете внимание, че това са клъстери на времеви серии от данни с дължина 492 месеца, макар че те са визуализирани като двумерни области. Белите области са области, в които налягане не е било еднородно. Клъстерите, разположени около полюсите, са с продълговата форма поради изкривяване, предизвикано от проектирането на сферичната повърхнина върху правоъгълник.



Фиг. 7-13. Клъстери от данни за налягане във времеви серии, намерени чрез използване на клъстерния алгоритъм, базиран на SNN-плътност

Използвайки SLP, учените определили времеви серии, наречени *климатични индекси*, които са полезни за анализа на различни феномени, свързани с климата на Земята. Например, аномалиите в климатичните индекси са свързани с аномално ниски или високи нива на превалявания или температура в различни части на света. Някои от клъстери, намерени чрез клъстеризацията, базирана на SNN-плътност, имат много силна връзка с някои от климатичните индекси, известни на учените, изследващи климат.

На Фиг. 7-14 са показана SNN-плътностна структура на данни, от които бяха построени клъстерите. Плътността е била нормализирана, за да бъде в диапазона от 0 до 1. Плътността на една времева серия може да изглежда като едно странно понятие, но тя измерва степента, в която една времева серия и нейните най-близки съседи имат едни и същи съседи. Тъй като всяка времева серия е асоциирана с конкретно местоположение, е възможно тези плътности да бъдат изобразени на една двумерна графика. Поради времева автокорелация, тези плътности формират смислени шаблони, т.е. е възможно визуално да бъдат идентифицирани клъстери от Фиг. 7-14.



Фиг. 7-14. Разпределение на SNN-плътност на данни за налягане, представени чрез времеви серии