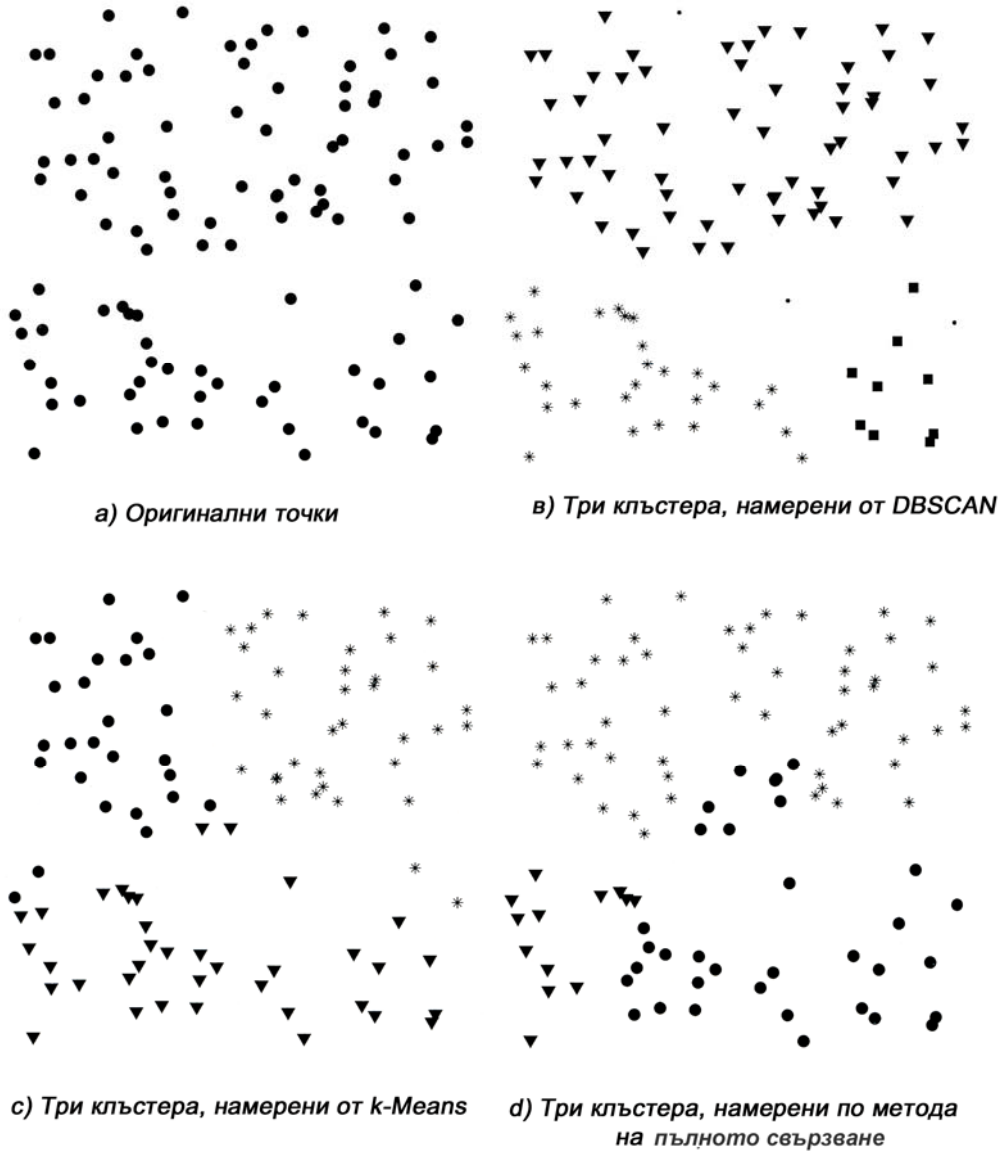
Лекция 8. Оценяване на клъстери

При класификация оценяването на получения класификационен модел е една неразривна част от процеса на създаването на класификационния модел и за таза цел са разработени няколко общопризнати мерки и процедури, като, например, класификационната точност, крос-валидацията и т.н. При клъстеризацията процедурата по оценяване на клъстери не е толкова добре разработена и не се използва задължително, като част от клъстерния анализ. Независимо от това, задачата по оценяване, или *валидация* (по-традиционен термин) *на клъстери* е една много важна част от клъстерния анализ и в тази лекция ще разгледаме най-често използвани и общоприети подходи към нейното решаване.



Фиг. 8-1. Клъстеризация на 100 равномерно разпределени случайни точки

Основната мотивация за необходимостта от оценяването на клъстери е фактът, че *всеки* алгоритъм за клъстеризацията е способен да намери клъстери в *произволно* множество

от данни, дори когато това множество *не съдържа* никакви естествени клъстери. Например, да разгледаме Фиг. 8-1, на която е представен резултат от клъстеризацията на 100 точки, случайно (равномерно) разпределени в един квадрат. Оригиналните точки са показани на Фис. 8-1(a), докато клъстерите, намерени от DBSCAN, K-means и чрез метода на пълното свързване са показани на Фиг. 8-1(b), 8-1(c) и 8-1(d). Тъй като DBSCAN е намерил три клъстера (чрез установяването на параметъра *Eps* на базата на разстояния до четири най-близки съседа), и за останалите два метода (използващи броят на клъстери като входния параметър), броят на клъстери е бил установен на три. Както се вижда, и в трите случая намерените клъстери не изглеждат „убедителни" (т.е. близки до естествени групи). Засегнатия в примера проблем е важи и при данни с по-голяма размерност, но в тези случаи тази „неубедителност" не може да бъде разбрана толкова лесно.

8.1. Класификация на подходи за оценяване на клъстери

Способността да определяме дали съществува някаква неслучайна структура в данни е само един от важните аспекти от задачата по валидацията на клъстери. Другитте важни аспекти от тази задача са:

1. Определяне на *клъстерната тенденция* в едно множество от данни, т.е. разпознаване дали някаква неслучайна структура действително съществува в данните.
2. Определяне на правилния брой на клъстери.
3. Оценяване доколко добре резултатите от клъстерния анализ „пасват" на изследваните данни *без използване* на никаква външна (за тези данни) информация.
4. Сравнение на резултатите от клъстерния анализ с резултатите, получени от други външни данни, например, предоставените отвън имена на класове.
5. Сравняване на два множества от клъстери с цел определяне кое от тях е по-добро.

Обърнете внимание, че проблемите 1, 2 и 3 не изискват използването на никаква външна информация – те се решават с използване на „ненаправлявани" техники (т.е. за обучение без учител - unsupervised), докато задача 4 изисква външната информация. Задачата 5 може да бъде решавана чрез използване както на ненаправлявани, така и направлявани (supervised) техники. Проблемите 3, 4 и 5 могат да бъдат още разделени в зависимост от това, дали искаме да оценяваме само отделни клъстери или цялата клъстеризация (т.е. всички клъстери).

При създаване на различни мерки за числово оценяване на различни аспекти на валидността на клъстери, трябва да отчитаме и следните моменти:

- Една мярка за оценяване на клъстери може да има доста ограничена област на приложение. Например, повечето от работите по създаване на мерки за оценяване на клъстерната тенденция са предназначени за работа с дву или тримерни пространствени данни.
- Необходима е ясна схема за интерпретация на всяка мярка. Например, ако получената стойност на мярката, оценяваща доколко добре етикетите (имената) на клъстери съответстват на предоставени отвън имена на класове, е 10, дали това означава, че съответствието е лошо, средно или добро?

- Ако една мярка е много сложна за прилагане или разбиране, много малко хора ще я използват.

Мерките за оценяване на клъстери (често наричани *клъстерни индекси*) традиционно се разделят на следните три типа:

1. *Ненаправлявани.* Това са мерки за оценяване на качество на клъстерната структура, не използващи никаква външна информация. Един пример за такава мярка е средно-квадратичната грешка. Често ненаправлявани мерки за оценяване на клъстери се разбиват на два класа:
   a. Мерки за *сплотеността (компактност, стегнатост*) на клъстера (cluster cohesion), оценяващи доколко са тясно свързани обекти, образуващи клъстера.
   b. Мерки за *изолираността на клъстера* (cluster separation), които определят доколко дадения клъстер е различен или добре отделен от други клъстери.

   Ненаправляваните мерки често се наричат *вътрешни индекси*, тъй като те използват само информацията, представена в конкретното множество от данни.
2. *Направлявани.* Тези мерки измерват до каква степен клъстерната структура, открита от клъстерния алгоритъм в даденото множество от данни, съответства (съвпада) с някоя външна структура. Един пример на такъв направляван индекс е ентропията, измерваща доколко добре етикетите на клъстери съвпадат с предоставените отвън етикетите на класове. Направляваните мерки често се наричат *външни* индекси, тъй като използват информация, която не присъства в даденото множество от данни.
3. *Относителни (relative)*. Тези мерки сравняват различни клъстеризации или клъстери. Относителната мярка за оценяване на клъстери е една ненаправлявана или направлявана мярка, която се ползва за целите на сравнение. Следователно, относителните мерки на практика не са отделен тип на мерки за оценяване на клъстери, а описват определен, специфичен начин за използване на такива мерки. Например, две различни клъстеризации, направени чрез алгоритъма k-Means, могат да бъдат сравнени както чрез използване на средно-квадратичната грешка, така и чрез ентропията.

8.2. Ненаправлявано оценяване на клъстери чрез използване на мерки за компактност и изолираност

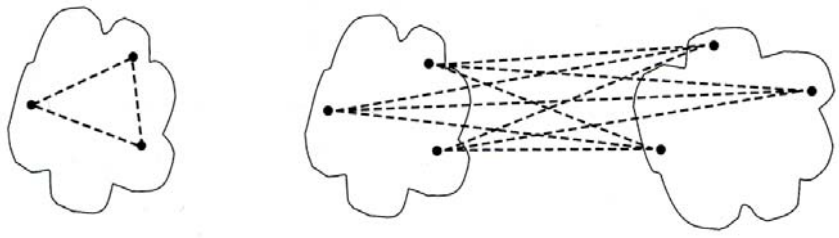
В общия случай валидността на една клъстеризация, състояща от K клъстера, може да се оцени като претеглената сума от оценките на отделни клъстери:

$$OverallValidity = \sum_{i=1}^{K} w_i * validity(C_i) \qquad (8.1)$$

Функцията *validity* може да измерва компактност, изолираност или да бъде някаква комбинация от тези две мерки. Теглата могат да варират в зависимост от използваната марка за валидност. В някои случаи те са просто 1 или равни на броя на обекти в клъстера, в други – отразяват някое по-сложно свойство, като например квадратен корен от компактността. Ако използваната функция за валидност на клъстера е компактността, то по-високите нейните стойности са по-добри; ако се използва изолираност – то по-добри са по-малките стойности.

8.2.1. Базиран на графи поглед на компактност и изолираност

За клъстери, разглеждани като графи, *компактността (cohesion)* на един граф може да бъде определена като сумата от теглата на всички връзки в графа на съседство (proximity), който съединява точки вътре в клъстера [1] (виж. Фиг. 8-2(а)). *Изолираността (separation)* на една двойка клъстери може да се измерва като сумата от теглата на връзки, съединяващи точки от един клъстер с точки от другия клъстер (виж Фиг. 8-2(в)).



а) Компактност в) Изолираност

Фиг. 8-2. Базиран на графи поглед на компактност и изолираност

Математически компактността и изолираността на един клъстер може да се изрази чрез формули (8.2) и (8.3):

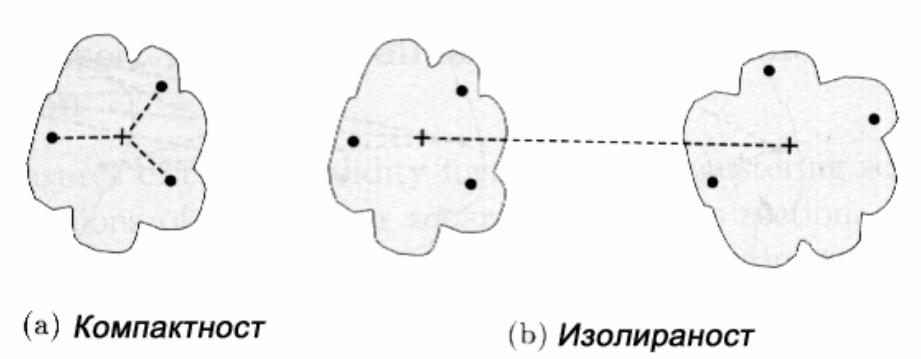
$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i \, \mathbf{y} \in C_i} proximity(\mathbf{x}, \mathbf{y}) \qquad (8.2)$$

$$separation(C_i, C_j) = \sum_{\mathbf{x} \in C_i \, \mathbf{y} \in C_j} proximity(\mathbf{x}, \mathbf{y}) \qquad (8.3)$$

Функцията на съседство (proximity) може да измерва сходството, разликата или някаква комбинация от тези две величини.

Базиран на прототипи поглед на компактност и изолираност

За клъстери, формирани чрез прототипи, компактността на един клъстер може да се дефинира като сумата от съседствата на всички точки от клъстера с неговия прототип (центроид или медоид). Аналогично, изолираността на една двойка клъстери може да се измерва чрез съседството между прототипите на тези клъстери (виж Фиг. 8-3, където прототипът на клъстера е означен като „+”).



Фиг. 8-3. Базиран на прототипи поглед на компактност и изолираност

[1] Напомням, че графът на съседство представя данни (точки) като възли и има връзка между всяка двойка от данни, на която се присвоява тегло, равно на степента на близост (отдалеченост) между възлите.

Математически компактността на един клъстер, базиран на прототипи, може да се изрази чрез формула (8.4), докато неговата изолираност – чрез две мерки, съответно (8.5) и (8.6), където $\mathbf{c_i}$ е прототипът (центроид) на клъстера C_i, а $\mathbf{c}$ е общия прототип (центроид) и на двата клъстера:

$$cohesion(C_i) = \sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c_i}) \qquad (8.4)$$

$$separation(C_i, C_j) = proximity(\mathbf{c_i}, \mathbf{c_j}) \qquad (8.5)$$

$$separation(C_i) = proximity(\mathbf{c_i}, \mathbf{c}) \qquad (8.6)$$

Двете мерки за изолираност отразяват факта, че, както ще видим след малко, изолираността на клъстерните прототипи от общия прототип е понякога директно свързана с изолираността на клъстерните прототипи един от другия. Обърнете внимание, че формулата (8.4) определя компактността на клъстера като неговата средно-квадратична грешка, ако в качеството на функцията на съседство е избрано Евклидовото разстояние.

8.2.2. Общи мерки за компактност и изолираност

Въведените по-горе дефиниции на клъстерната компактност и изолираност представляват прости и добре-дефинирани мерки за оценка на клъстери, които могат да бъдат комбинирани в една обща мярка за валидността на клъстера, използвайки претеглената сума съгласно формула (8.1). Обаче, за тази цел трябва да дефинираме теглата. Съществуват различни начини за дефиниране на теглата, макар, обикновено, те отразяват някаква мярка за размер на клъстера.

В таблица 8-1 са представени няколко мерки за валидността на клъстери, базирани на клъстерните компактност и изолираност.

Име	Клъстерна мярка	Клъстерно тегло	Тип
I_1	$\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_i} proximity(\mathbf{x}, \mathbf{y})$	$\frac{1}{m_i}$	Компактност, базирана на графи
I_2	$\sum_{\mathbf{x} \in C_i} proximity(\mathbf{x}, \mathbf{c_i})$	1	Компактност, базирана на прототипи
E_1	$proximity(\mathbf{c_i}, \mathbf{c})$	m_i	Изолираност, базирана на графи
Γ_1	$\sum_{j=1}^{k} \sum_{j \neq i} \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_i} proximity(\mathbf{x}, \mathbf{y})$	$\frac{1}{\sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_i} proximity(\mathbf{x}, \mathbf{y})}$	Изолираност и компактност, базирана на графи

Таблица 8.1 Общи мерки за оценяване на валидността на клъстери

Мярка $\mathbf{I_1}$ е общата мярка за компактност, изразена в термините на взаимното съседство между обектите в клъстера, делено върху размера на клъстера. Мярка $\mathbf{I_2}$ е общата мярка за компактност, базирана върху сумата на съседствата между обекти в клъстера и клъстерния центроид. Мярка $\mathbf{E_1}$ е общата мярка за компактност, изразена като съседството на клъстерния центроид с общия центроид на клъстеризацията, умножено по броя на обекти в клъстера. Мярка $\mathbf{\Gamma_1}$ е общата мярка за компактност и изолираност, която представлява сумата на взаимното съседство между всички обектите от различни клъстери (т.е. общото тегло на всички връзки от графа на съседството, които трябва да

бъдат прекъснати, за да бъде изолиран един клъстер от всички останали клъстери), делена върху сумата на взаимното съседство на всички обекти в самия клъстер.

Трябва да се отбележи, че всяка ненаправлявана мярка за валидността на клъстера може потенциално да бъде използвана в един клъстерен алгоритъм като функция за годност, направляваща процеса на клъстеризацията. Т.е. всяка точка се назначава към клъстера, който дава най-добрата стойност на използваната в алгоритъма функция на годност. Например, мярка за валидността на клъстери $\mathbf{I_2}$ съответства на традиционен алгоритъм k-Means и произвежда клъстери с добри стойности на средно-квадратичната грешка (SSE). Другите мерки произвеждат клъстери, които не са толкова добри по отношение на SSE, но са по-оптимални по отношение на избраната мярка за валидността на клъстери.

8.2.3. Връзка между компактността, базирана на прототипи, и компактността, базирана на графи

Макар че мерките, базирани върху различни погледи на клъстер (като граф или като прототип) изглеждат на пръв поглед различни, при избора на някои конкретни мерки за оценяване на съседство на обекти те стават еквивалентни. Например, при избора на Евклидовото разстояние в качество на мярка за съседство, мярката $\mathbf{I_2}$ (която ще съответства на SSE) е еквивалентна на мярка $\mathbf{I_1}$:

$$SSE(C_i) = \sum_{\mathbf{x} \in C_i} dist(\mathbf{c_i}, \mathbf{x})^2 = \frac{1}{2m_i} \sum_{\mathbf{x} \in C_i} \sum_{\mathbf{y} \in C_i} dist(\mathbf{x}, \mathbf{y})^2 \qquad (8.7)$$

8.2.4. Два подхода към измерване на изолираността, базирана на прототипи

Когато съседството се измерва чрез Евклидовото разстояние, традиционната мярка за изолираност (различие) между клъстери се превръща в междуклъстерната сума от квадрати ($SSB(C_i)$), която се изчислява като сумата от квадрати на разстояния от центроида $\mathbf{c_i}$ до общия центроид на всички точки $\mathbf{c}$. Чрез сумирането на $SSB(C_i)$ по всички клъстери получаваме общата междуклъстерната сума SSB:

$$SSB = \sum_{i=1}^{K} SSB(C_i) = \sum_{i=1}^{K} m_i dist(\mathbf{c}_i, \mathbf{c})^2 \qquad (8.8)$$

Лесно е да се покаже, че SSB е пряко свързано с разстояния между центроиди по двойки. В частност, ако размерите на клъстери са еднакви, т.е. $m_i = m/K$, то тази връзка приема простата форма, задавана от формула (8.9):

$$SSB = \frac{1}{2K} \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m}{K} dist(\mathbf{c}_i, \mathbf{c}_j)^2 \qquad (8.9)$$

Точно този тип на еквивалентността мотивира дефинирането на изолираността на прототипи по два начина – чрез формулите (8.5) и (8.6).

8.2.5. Връзка между компактност и изолираност

В някои случаи съществува строга зависимост между компактността и изолираността. Така е възможно да се покаже, че сумата на квадрати на вътрешноклъстерни (SSE) и външноклъстерни (SSB) грешки е една постоянна величина, която е равна общата сума

на квадрати (*TSS*), която е сума от квадрати на разстояния от всяка точка до общия център (центроид) на всички данни. Важността на този резултат се състои в това, че минимизацията на *SSE* (компактността) е еквивалентна на максимизацията на *SSB* (изолираността).

Да докажем това утвърждение, предполагайки за опростяване на нотацията, че данните са едномерни, т.е. $dist(x,y) = (x - y)^2$. Освен това, ще използваме и факт, че $\sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)(c-c_i) = 0$:

$$TSS = \sum_{i=1}^{K}\sum_{x\in C_i}(x-c)^2 = \sum_{i=1}^{K}\sum_{x\in C_i}((x-c_i)-(c-c_i))^2 =$$

$$= \sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)^2 - 2\sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)(c-c_i) + \sum_{i=1}^{K}\sum_{x\in C_i}(c-c_i)^2 =$$

$$= \sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)^2 + \sum_{i=1}^{K}\sum_{x\in C_i}(c-c_i)^2 =$$

$$= \sum_{i=1}^{K}\sum_{x\in C_i}(x-c_i)^2 + \sum_{i=1}^{K}|C_i|(c-c_i)^2 = SSE + SSB$$

8.2.6. Оценяване на отделни клъстери и обекти

До сега ние сме се фокусирали върху използване на мерки за изолацията и компактността за общото оценяване на валидност на някоя група от клъстери. Повечето от тези мерки за валидността на клъстери могат да се ползват и за оценка на отделни клъстери или обекти. Например, можем да ранжираме отделни клъстери в съответствие с техните стойности на валидността, т.е. клъстерната компактност или изолираност. Един клъстер, който има висока стойност на компактността може да се смята като по-добър от този с по-малка стойност. Тази информация може да се използва за подобряване на качеството на клъстеризацията. Ако, например, един клъстер не е много компактен, можем да го разделим на няколко подклъстера. От другата страна, ако два клъстера са относително компактни, но не са добре изолирани, можем да опитаме да ги слеем в един клъстер.

Ние също така можем и да оценяваме обектите в клъстера в термините на техния принос към общата компактност или изолираност на съответния клъстер. Обектите, които допринасят повече за компактността и изолираността, са разположени „по-навътре" в клъстера. Обратно, обектите с малък принос към тези показатели са по-вероятно да са разположени около „края" на клъстера. В следващия раздел ще разгледаме една мярка за оценка на клъстера, която използва подходът, базиран на тези идеи, за оценяване на точки, клъстери и целия набор от клъстери.

Коефициент на силуета

Един популярен метод, използващ така наречени „коефициенти на силуета", комбинира както клъстерната компактност, така и изолираност. По-долу е приведен алгоритмът за изчисляване на коефициента на силуета за всяка отделна точка, като процесът се състои от три стъпки. В алгоритъма в качеството на мярка за съседството се използва разстояние, но същия подход може да бъде приложен и при използването на сходството:

Стъпка 1. За всеки i-ти обект изчисли неговото средно разстояние a_i до всички други обекти от същия клъстер.

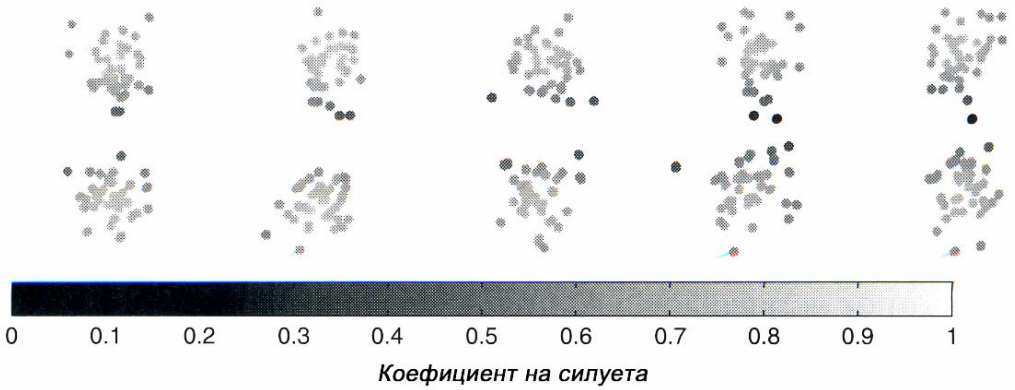
Стъпка 2. За всеки i-ти обект и всеки клъстер, не съдържащ този обект, изчисли средното разстояние от обекта до всички обекти от дадения клъстер. Намери минимум на тази величина и й обозначи чрез b_i.

Стъпка 3. За всеки i-ти обект изчисли неговия коефициент на силует s_i по формула:

$$s_i = \frac{(b_i - a_i)}{\max\{a_i, b_i\}}$$

Стойността на коефициента на силуета може да варира от -1 до 1. Отрицателната стойност е нежелателна, тъй като това отговаря на случая, когато a_i, средното разстояние до точки от същия клъстер, е по-голямо от b_i – минималното средно разстояние до точки от други клъстер. Желателно е коефициентът на силуета да е положителен ($a_i < b_i$) и a_i да бъде колкото може по-близо до нула, тъй като коефициентът приема своята максимално стойност 1 при $a_i = 0$.

Коефициентът на силуета на един клъстер се изчислява като средното на коефициентите на силуета на всички точки от този клъстер. Общият коефициент на силуета на цялата клъстеризация се получава чрез осредняване на коефициентите на силуета на всички клъстери. Фиг. 8-4 показва графично коефициенти на силуета за точки в 10 клъстера. По-тъмен цвят отговаря на по-малки коефициенти на силуета.



Фиг. 8-4. Коефициенти на силуета за точки от десет клъстера

8.3. Ненаправлявано оценяване на клъстери с използване на матрицата на сходство

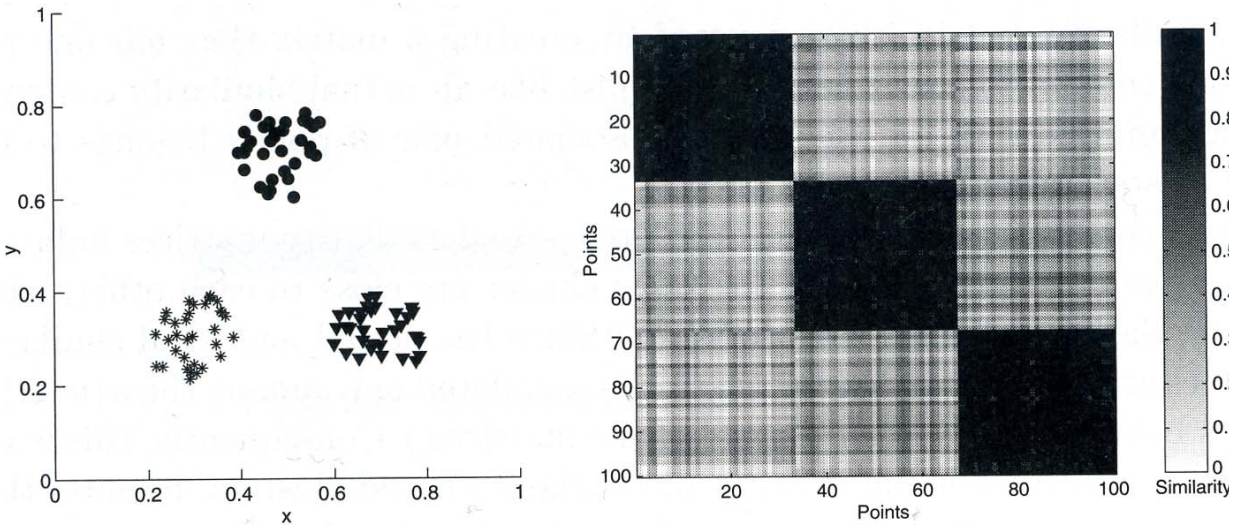
8.3.1. Измерване на валидността на клъстери чрез корелация

Ако от зададеното множество данни изчислим матрицата на сходство и получим етикетите на клъстери от проведения клъстерен анализ, то можем да оценим качеството на получената клъстеризация чрез анализ на корелацията между изчислената матрица на сходството и една нейна идеализирана версия, получена само на базата на етикетите на клъстери (с малки промени, описаният по-долу подход може да бъде приложен и за матриците на съседство, но за да опростим описанието на идеята, ще се ограничим само с матрицата на сходство).

Един идеален клъстер е този, който съдържа само точки, които имат сходството 1 с всички точки в клъстера и сходството 0 със всички точки от други клъстери. Следователно, ако сортираме редове и колонки на матрицата на сходство по такъв начин, че всички обекти, принадлежащи към един и същ клъстер са събрани заедно, то идеалната матрица на сходство ще има блок-диагоналната структура. С други думи, сходството ще е 1 вътре в тези блокове на матрицата, в които елементите ще представляват вътреклъстерното сходство, и 0 извън тях. Идеалната матрица на сходство се конструира чрез създаване на матрицата, която има по един ред и по-една колонка за всяка точка данни – по същия начин както и актуалната матрица на сходство – и назначаване на 1 на всеки елемент, ако асоцииранатa с него двойка от точки принадлежи към един и същ клъстер, или 0 – в противен случай.

Високата корелация между идеалната и истинската матрици на сходство показва, че точките, принадлежащи към един и същ клъстер, са близко една до друга, докато низката корелация означава обратното. (Тъй като действителната и идеалната матрици на сходство са симетрични, корелацията се изчислява само между *n(n-1)/2* елемента под или над диагоналата на матриците.) От тук следва, че тази оценка не е добра мярка за клъстери, формирани от алгоритми, базиращи на плътността, тъй като те нямат сферична форма и могат тясно да се преплитат с други клъстери.

За да илюстрираме описания по-горе подход да изчислим корелацията между идеалната и действителната матрици на сходство за клъстери, формирани от алгоритъма k-Means, показани на Фиг. 8-1 (c) (случайни данни) и Фиг. 8-5(a) (данни с три добре-отделени клъстера). Изчислените корелации са 0.5810 и 0.9235 съответно, което отразява очаквания резултат, че клъстерите, намери от k-Means в случайните данни, са по-лоши от тези, намерени в данни с добре отделени клъстери.



а) Добре отделени клъстери

в) Матрица на сходство, сортирана по етикети на клъстери, определяни от k-Means

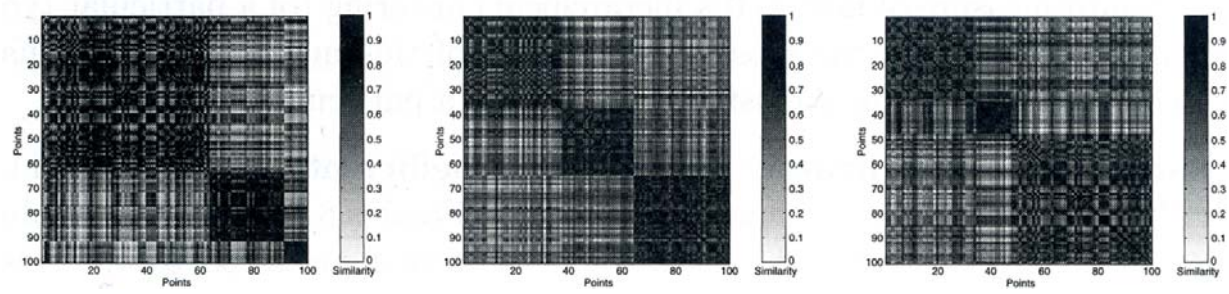
Фиг. 8-4. Коефициенти на силуета за точки от десет клъстера

8.3.2. Визуално оценяване на клъстеризация чрез използване на матриците на сходство

Описаната по-горе техника предлага един по-общ, качествен подход към оценяване на множество от клъстери: подреждане на матрицата на сходство по отношение на клъстерните етикети и след това – нейната визуализация. Теоретично, ако имаме добре отделение клъстери, показваните шаблони трябва да бъдат близки до блок-диагоналната форма. Ако не, те трябва да показват зависимости между клъстери.

В качеството на пример, да разгледаме точки от Фиг. 8-5(а), които оформят три добре отделени клъстера. По тази причина те лесно могат да бъдат намерени от K-means. Изолираността на намерените клъстери се илюстрира чрез пренареждане на матрицата на сходство, показана на Фиг. 8-5(в). (Сходствата са били получени от разстояния чрез прилагането на формулата $s = 1 - (d - min\_d) / (max\_d - min\_d)$). Фиг. 8-6 показва пренаредените матрици на сходство за клъстери, намерени в множеството от случайни данни от Фиг. 8-1 чрез прилагане на алгоритмите DBSCAN, k-Means и пълното свързване.

Добре отделените клъстери от Фиг. 8-5 показват наличието на силно изразени блок-диагонални шаблони в пренаредената матрицата на сходство. Обаче, по-слабо изразените блок-диагонални шаблони присъстват и в пренаредените матрици на сходство от Фиг. 8-6, намерени от алгоритмите DBSCAN, k-Means и пълното свързване в случайни данни. Точно като хората са способни да намират шаблони в облаци, алгоритмите за извличане на закономерности от данни могат да намират шаблони и в случайни данни. Докато намирането на някакви шаблони (изображения) в облаци е развлечение, намирането на някакви закономерности в случайни данни (шум) е безсмислено и дори объркващо.



а) Матрица на сходство, сортирана по клъстерни етикети на DBSCAN

в) Матрица на сходство, сортирана по клъстерни етикети на k-Means

с) Матрица на сходство, сортирана по клъстерни етикети на метода за пълното свързване

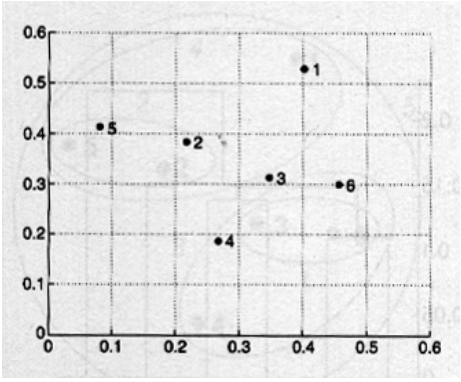
Фиг. 8-6. Матрици на сходство за клъстери, извлечени от случайни данни

Предлаганият подход е изглежда безнадеждно скъп за прилагане към големи по обем множества от данни, тъй като изчисляване на матрицата на съседство изисква време от порядъка на $O(m^2)$, където m е броят на обекти. Обаче методът може да бъде прилаган към извадки от данни – можем да вземем някаква случайна извадка от точки от всеки клъстер, да изчислим матрицата на сходство между тези точки и след това да визуализираме резултата. В някои случаи за получаване на адекватното представяне на всички клъстери се налага да увеличаваме извадката (oversampling) на малки клъстери и да намаляваме (undersampling) размер на извадката за по-големи клъстери.

8.4. Ненаправлявана оценка на йерархична клъстеризация

Предишните подходи към оценяване на клъстери са предназначени за нейерархичната клъстеризация. В този раздел ще разгледаме една популярна марка за оценка на качество на йерархичната клъстеризация – така наречена кофенетична (*cophenetic*) корелация. *Кофенетичното разстояние* между два обекта е стойността на съседството, при която един агломеративен алгоритъм за йерархична клъстеризация за първи път групира заедно тези обекти в един и същ клъстер. Например, ако в някой от етапите на процеса на агломеративна клъстеризация най-малкото разстояние между двата клъстера, подлежащи на сливане, е равно на 0.1, то всички точки в един от тези клъстери имат кофенетичното разстояние равно на 0.1 до всяка от точките, принадлежащи към другия клъстер. Кофенетичното разстояние между всички двойки обекти се отразява в матрицата на кофенетичното разстояние. Кофенетичното разстояние между една и съща двойка точки е различно за различни алгоритми за агломеративна клъстеризация.

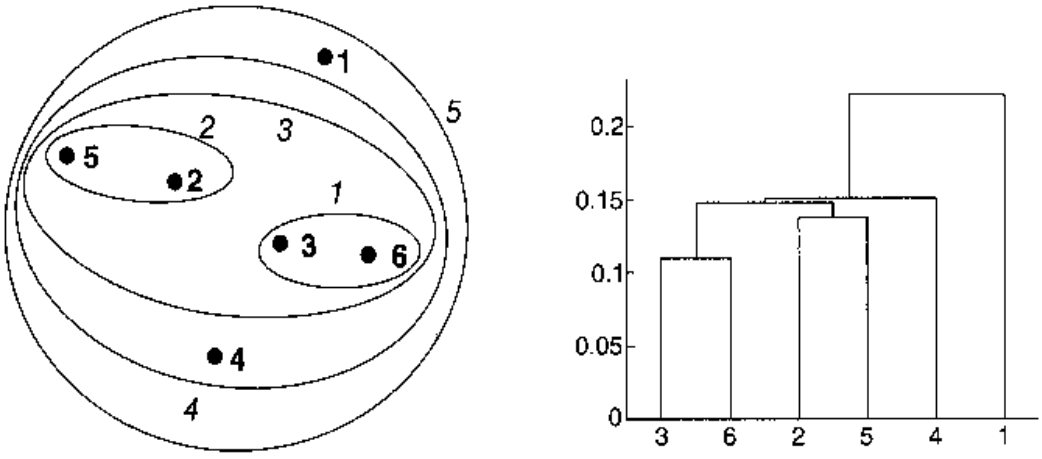
В качеството на пример, да разгледаме множеството от 6 двумерни точки, показани на Фиг. 8-7. Матрицата на Евклидовото разстояние между тях е показана в Таблица 8-2:



Фиг. 8-7. Множество от 6 двумерни точки

Точка	P1	P2	P3	P4	P5	P6
P1	0.00	0.24	0.22	0.37	0.34	0.23
P2	0.24	0.00	0.15	0.20	0.14	0.25
P3	0.22	0.15	0.00	0.15	0.28	0.11
P4	0.37	0.20	0.15	0.00	0.29	0.22
P5	0.34	0.14	0.28	0.29	0.00	0.39
P6	0.23	0.25	0.11	0.22	0.29	0.00

Таблица 8-2. Матрица на Евклидови разстояния между шестте точки



Фиг. 8-8. Клъстеризация на 6 точки от Фиг. 8-7 по метода на единичното свързване

Ако за клъстеризация на тези точки ще се използва агломеративният алгоритъм, базиран на метода на единичното свързване (виж Фиг. 8-8), то матрицата на кофенетичното разстояние между тези точки ще изглежда по следния начин (виж Таблица 8-3)

Точка	**P1**	**P2**	**P3**	**P4**	**P5**	**P6**
P1	0.00	0.22	0.22	0.22	0.22	0.22
P2	0.22	0.00	0.15	0.15	0.14	0.15
P3	0.22	0.15	0.00	0.15	0.15	0.11
P4	0.22	0.15	0.15	0.00	0.15	0.15
P5	0.22	0.14	0.15	0.15	0.00	0.15
P6	0.22	0.15	0.11	0.15	0.15	0.00

Таблица 8-3. Матрица на кофенетични разстояния между шестте точки

Коефициентът на кофенетична корелация (CPCC) е корелацията между елементите матрицата на кофенетични разстояния и оригиналната матрицата на различия. Той е една стандартна мярка за оценяване, доколко добре получената йерархична клъстеризация отговаря на данните. Коефициентът се изчислява по следната формула:

$$CPPC = \frac{\sum_{i<j}(dist(x_i,x_j) - \overline{dist(x)})(CP\_dist(x_i,x_j) - \overline{CP\_dist(x)})}{\sqrt{\sum_{i<j}(dist(x_i,x_j) - \overline{dist(x)})^2 \sum_{i<j}(CP\_dist(x_i,x_j) - \overline{CP\_dist(x)})^2}} \quad (8.10)$$

Където dist(.,.) и CP_dist(.,.) са, съответно, функциите за разстояние и кофенетичното разстояние между точките, а $\overline{dist(x)}$ и $\overline{CP\_dist(x)}$ са съответните средните разстояния. Най-често коефициентът на кофенетичната корелация се използва за определяне, кой тип от йерархичната клъстеризация е по-добре пасва на конкретния набор от данни.

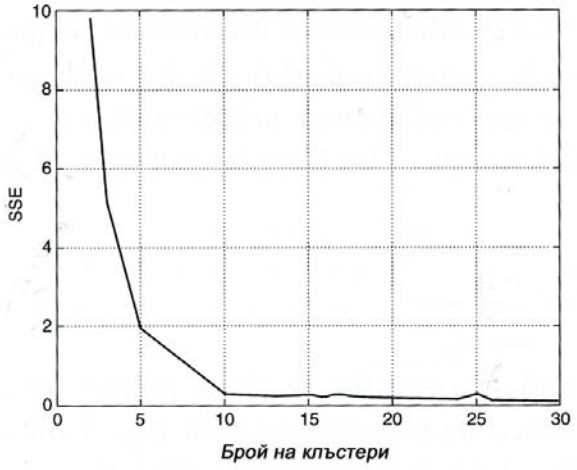
В таблица 8-4 са показани изчислените стойности на този коефициент за описаните по-горе 6 точки, получени при използване на 4 различни техники за агломеративна клъстеризация. Както се вижда от таблицата, йерархичната клъстеризация, получена чрез техника за единичното свързване, изглежда пасва по-лошо на данните, отколко при използване на метода на пълното свързване или груповото осредняване.

Метод	**CPCC**
Единично свързване	0.44
Пълно свързване	0.63
Групово осредняване	0.66

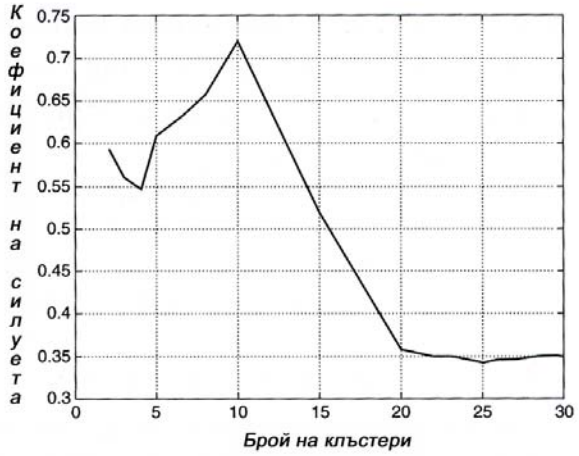
Таблица 8-4. Коефициент на кофенетичната корелация за данни от Таблица 8-2 за 3 различни метода за агломеративната клъстеризация

8.5. Определяне на коректен брой клъстери

За приблизително определяне на коректен или естествен брой клъстери в данни могат да бъдат използвани различни ненаправлявани мерки за валидността на клъстеризация.



Фиг. 8-9. Графика на SSE срещу броя на клъстери за данни от Фиг. 8-4



Фиг. 8-10. Графика на средния коефициент на силуета срещу броя на клъстери за данни от Фиг. 8-4

Да разгледаме пак данните, показани на Фиг. 8-4, които имат естествени 10 клъстера. Да изберем за клъстеризацията разполовяващия (bisecting) вариант на алгоритъма k-Means. На Фиг. 8-9 е показан графикът на промяна на стойността на средноквадратичната грешка (SSE) в зависимост от броя на клъстери, а на Фиг. 8-10 – зависимостта на средния коефициент на силуета от броя на клъстери за същите данни. Първият график показва наличието на ясно изразен прелом в кривата на SSE, а вторият – ясно изразен пик в коефициента на силуета, когато броят на клъстери е равен на 10.

Следователно, можем да определим естествения брой на клъстери в данни чрез търсене на такъв брой клъстери, при които съществуват преломи, пикове или хлътвания в графици на използваните оценъчните функции за качеството на клъстеризацията. Естествено, че този подход не винаги дава добри резултата. Клъстерите могат да бъдат в много по-голяма степен преплетени или препокриващите се, отколко тези, използвани в примера. Освен това, те могат да бъдат вложени. В действителност, дори в използвания пример те са донякъде вложени – в данните има 5 двойки клъстери, тъй като клъстерите са по-близки по вертикала, отколко по хоризонтала. Този факт е отбелязан с прелом в кривата на SSE, но не е толкова ясно изразен в кривата на коефициента на силуета. С други думи, предлаганият подход може да помогне в намирането на естествения брой клъстери в данни, но неговите резултати се нуждаят от внимателната интерпретация.

8.6. Клъстерна тенденция

Един очевиден начин да се определи, дали данните съдържат някакви клъстери, е да се направи опит да ги клъстеризираме. Обаче, почти всички клъстерни алгоритми ще намерят някакви клъстери в данни. Следователно, за да кажем, че в данните се съдържат някакви клъстери, трябва да оценим намерените клъстери и да преценим, че поне някои от тях са с добро качество. Обаче, този подход не отчита, че истинските клъстери могат да имат форма, отлична от тази, търсена от прилагания клъстерен

алгоритъм. За да се справим с тази допълнителна трудност можем да използваме няколко различни клъстерни алгоритми и отново да оценим качеството на получените резултати. Ако клъстерите навсякъде са с лошо качество, това може да бъде индикация, че данните просто не съдържат естествени клъстери.

Алтернативният подход, и точно върху това се акцентира в мерките за клъстерната тенденция, е да се опитаме да определим дали данните съдържат някакви клъстери без изобщо да опитваме да ги клъстеризираме. Най-често използваният подход, особено за данни в Евклидовото пространство, се състои в прилагане на статистически тестове за оценка на пространствена еднородност. За съжаление изборът на коректния модел, оценяване на неговите параметри и на статистическата значимост на хипотезата, че данните не са случайни, е доста сложна задача. Но не зависимо от това, за нейното решаване са разработени няколко подхода, като повечето от тях са за точки от Евклидовото пространство с малка размерност.

Като пример, ще разгледаме така наречена статистиката на Хопкинс (Hopkins). При този метод се генерират p точки, които са случайно разпределени в пространството от данни. Освен това от реалните данни се избират по случаен начин също p реални точки. За всяко от получените два множества от точки се изчислява разстояние до най-близките съседи от оригиналното множество от данни. Нека чрез w_i да означим разстоянието до най-близкия съсед от i-та точка от подмножеството от реални данни, а чрез u_i – разстоянието до най-близкия съсед от i-та точка от подмножеството от изкуствено генерирани случайни данни. Статистиката на Хопкинс H се изчислява по следната формула:

$$H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} w_i + \sum_{i=1}^{p} u_i} \qquad (8.11)$$

Ако случайно генерираните точки и избраното подмножество от реалните точки имат приблизително едно и също разстояние до най-близките съседи, то H ще бъде близко до 0.5. Стойностите на H близки до 0 или 1 показват, съответно, че данните са или силно клъстеризирани, или равномерно разпределени в пространството на данни. Като пример за използване на тази статистика, да изчислим нейната стойност за данни от Фиг. 8-1 (без ясно изразени клъстери) и от Фиг. 8-4 (с 3 добре отделени клъстери). Ще изчислим статистиката при 100 случайни опита за формиране на множества от данни с размер $p = 20$. За първите данни стойността на H = 0.56 със стандартното отклонение 0.03. За вторите данни H = 0.095 със стандартното отклонение 0.006.

8.7. Направлявани мерки за оценка на валидността на клъстери

Обикновено наличието на външна информация за данни се свежда до предоставените отвън етикети за класовете на обектите от данни. В такива случаи прилаганите процедури се свеждат до установяване на съответствие между тези външни етикети и етикетите на клъстери, получени след прилагане на някой клъстеризационен алгоритъм. Защо се прави това? Защо изобщо трябва да се прави клъстерния анализ, когато е известно разбиването на данни по класове? Мотивацията за подобен анализ е или сравнение на клъстерните техники с „естествено” групиране на данни или да прецени,

до каква степен „ръчният" режим на класификацията може да бъде автоматизиран чрез използване на клъстерния анализ.

Ще разгледаме два типа подходи. Първият набор от техники използва мерки, взети от задачата по класификация, такива като ентропия, чистота и F-мярка. Тези мерки оценяват, до каква степен един клъстер съдържа обекти, принадлежащи към един единствен клас. Втората група методи е свързана с мерки за сходство, прилагани към двоични данни, такива като, например, Жакардов коефициент. Тези подходи измерват до каква степен два обекта, принадлежащи към един и същ клас, са групирани в един и същ клъстер и обратно. За удобство, ще наричаме тези два типа мерки *класификационно ориентирани* и *ориентирани на сходство* съответно.

8.7.1. Класификационно ориентирани мерки за оценка на валидността на клъстери

За оценяване на поведение на един класификационен модел е прието да се използват такива мерки като ентропия, чистота, прецизност, припомняне и F-мярка. В случай на класификацията ние измерваме степента, в която предсказания етикет на класа съответства на актуалния етикет на класа, но и в случая на клъстеризацията нищо фундаментално не се променя в използването на указаните мерки, освен че вместо предсказаните етикети на класа ще се ползват етикетите на клъстери.

Ентропия: степента, в която един клъстер се състои от обекти, принадлежащи само към един клас. За всеки клъстер отначало се изчислява разпределение на данни по класове, т.е. за клъстер i се изчислява p_{ij} – вероятността, че един обект от клъстера i принадлежи към класа j: $p_{ij} = m_{ij} / m_i$, където m_i е броят на обекти в клъстера i , а m_{ij} е броят на обекти от този клъстер, принадлежащи към класа j. Използвайки изчисленото разпределение по класове, ентропията на всеки клъстер i се изчислява по стандартната формула $Entropy_i = -\sum_{j=1}^{L} p_{ij} \log_2 p_{ij}$, където L е броят на класове. Общата ентропия на клъстеризацията се изчислява като сума от ентропиите на всеки клъстер, претеглена по размера на всеки клъстер, т.е. $Entropy = \sum_{i=1}^{K} \frac{m_i}{m} Entropy_i$, където m е общия брой на обекти.

Чистота (Purity): това е друга мярка, оценяваща степента, в която един клъстер се състои от обекти, принадлежащи само към един клас. Използвайки вече въведените обозначения, чистотата на един клъстер е

$Purity_i = \max_j \{p_{ij}\}$, а общата чистота на клъстеризацията е $Purity = \sum_{i=1}^{K} \frac{m_i}{m} Purity_i$

Прецизност (Precision): пропорцията на обекти от един клъстер, принадлежащи към някой конкретен клас. С други думи, прецизността на клъстер i по отношение на класа j е $precision(i, j) = p_{ij}$.

Припомняне (Recall): степента, в която един клъстер съдържа всички обекти от указания клас. Припомнянето на клъстер i по отношение на класа j е $recall(i, j) = m_{ij}/m_j$.

F-мярка: определена комбинация от прецизността и припомнянето, която измерва степента, в която един клъстер съдържа *само* обекти дадения клас и *всички* обекти от този клас. F-мярката на клъстер i по отношение на класа j е

$$F(i, j) = \frac{2\,precision(i, j) \times recall(i, j)}{precision(i, j) + recall(i, j)} \quad (8.12)$$

В качеството на примера, илюстриращ тези мерки, да разгледаме Таблица 8.5, отразяваща резултати от клъстеризацията на 3204 статии във вестника *Los Angeles Times* с помощта на алгоритъма k-Means, използващ като мярка за сходство косинус между векторното представяне на текстове. Тези текстове принадлежат към различни класове (тематики): развлечение, финанси, международни новини, времето, национални новини и спорт. Таблицата показва клъстеризацията в 6 клъстера. Първата колонка указва номер (индекс) на клъстера, докато останалите 6 колонки образуват матрицата на несъответствия (confusion matrix), т.е. тези колонки показват, как документи от различни класове са разпределени между клъстерите. Последната колонка показва стойностите на ентропията и чистотата за всеки от клъстерите.

Клъстер	Развлече-ния	Финанси	Чуждестр. новини	Времето	Новини от страната	Спорт	Ентропия	Чистота
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Общо	354	555	341	943	273	738	1.1450	0.7203

Таблица 8.5. Резултатите от k-Means клъстеризация на множеството документи от Los Angeles Times

В идеалния случай всеки клъстер трябва да съдържа документи само от един клас. В реалния случай всеки клъстер съдържа документи от няколко класа. Обаче, много от клъстери съдържат документи предимно от един клас. В частност, клъстер 3, който съдържа основно документи, свързани със спорта, е много добър както в термините на ентропия, така и в термините на чистотата. Чистотата и ентропията на други клъстери не са толкова добри, но обикновено, могат да бъдат значително подобрени, ако данните се разбивани на по-голям брой клъстери.

Прецизността, припомнянето и F-мярката могат лесно да бъдат изчислени за всеки клъстер. Например, за клъстер 1 и клас „Времето" прецизността е 506/677 = 0.75, а припомнянето е 506/943 = 0.25. Следователно F-мярка е 0.39. Същата мярка за клъстер 3 и „Спорт" е 0.95.

8.7.2. Ориентирани на сходство мерки за валидността на клъстери

Всички мерки, които ще се обсъждат в този раздел, се базират на предположението, че всяка двойка обекти, принадлежащи към един и същ клъстер, трябва да принадлежат и към един и същ клас. Този подход към оценяване на валидността на клъстери може да се разглежда като сравнение на две матрици: 1) *идеалната матрица на сходство на клъстери,* в която стойността на елемент *ij* е равна на 1, ако и двата обекта *i* и *j* принадлежат към един и същ клъстер или 0, в противен случай, и 2) *идеалната матрица на сходство на класове*, в която стойността на елемент *ij* е равна на 1, ако и двата обекта *i* и *j* принадлежат към един и същ клас или 0, в противен случай. Както и по-рано, в качеството на мярка за валидността на клъстери се използва корелацията между тези две матрици. Тази мярка е известна в литература като *Г-статистика*.

В качеството на пример да разгледаме 5 точки $p_1, \ldots, p_5$, които са разпределени в два клъстера $C_1 = \{p_1, p_2, p_3\}$ и $C_2 = \{p_4, p_5\}$ и два класа $L_1 = \{p_1, p_2\}$ и $L_2 = \{p_3, p_4, p_5\}$. Идеалните матрици на сходство за клъстери и класове са показани в Таблица 8-6 и 8-7 съответно. Корелацията на стойностите от тези две матрици са 0.359.

Точка	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Таблица 8-6. Идеалната матрица на сходство на клъстери

Точка	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

Таблица 8-7. Идеалната матрица на сходство на класове

В общия случай за оценката на валидността на клъстери може да се използва всяка мярка за сходство, базирана на двоични вектори (например, използваните в примера матрици могат да се превърнат в двоични вектори чрез обединяването на редове). Ще повторим дефинициите на тези мерки, модифицирайки ги, за да се нагласят към дадения контекст. По-точно, трябва да изчислим следните 4 стойности за всички двойки различни обекти (това е $m(m - 1)/2$ двойки, където m е броят на обектите).

f_{00} = броят на двойки на обекти от различни класове, принадлежащи към различни клъстери
f_{01} = броят на двойки на обекти от различни класове, принадлежащи към един и същ клъстер
f_{10} = броят на двойки на обекти от един и същ клас и принадлежащи към различни клъстери
f_{11} = броят на двойки на обекти от един и същ клас и принадлежащи към един и същ клъстер

Най-често използваните мерки са коефициентът на простото съвпадение (още известен като статистиката на Rand) и Жакардов коефициент:

$$Rand = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \quad (8.13) \qquad Jaccard = \frac{f_{11}}{f_{01} + f_{10} + f_{11}} \quad (8.14)$$

Да изчислим тези статистики за примера, показан с таблиците 8.6 и 8.7. Те се преобразуват в следните двоични вектори:

p_{21}	p_{31}	p_{41}	p_{51}	p_{32}	p_{42}	p_{52}	p_{43}	p_{53}	p_{54}
1	1	0	0	1	0	0	0	0	1
1	0	0	0	0	0	0	1	1	1

Отбелязвайки, че $f_{00} = 4$, $f_{01} = 2$, $f_{10} = 2$ и $f_{11} = 2$, получаваме, че статистиката на Rand = 0.6, а Жакардов коефициент е 0.33.

8.7.3. Оценка на валидността на клъстери при йерархична клъстеризация

Оценката на валидността на йерархичната клъстеризация е по-сложна от нейерархичната по няколко причини, сред които е, че често в данни не съществува никаква предварително известна йерархична структура (от класове). По тази причина ще разгледаме един подход към оценяване на йерархичната клъстеризация в термини на едно плоско множество от класове, което се смятаме за предварително известно.

Ключовата идея на подхода е да преценим, дали йерархичната клъстеризация съдържа за всеки клас поне един клъстер, който е относително еднороден и съдържа повечето от обектите от този клас. За тази цел можем за всеки клас и всеки клъстер в йерархията да изчислим F-мярката. След това за всеки клас избираме максималната стойност на F-мярката, свързана с всеки клъстер. На края изчисляваме общата стойност на F-мярката за цялата йерархична клъстеризация като претегленото средно от всички по-класови F-мерки, където теглата се базират върху размерите на класове. С други думи, тази йерархичната F-мярка се изчислява по следната формула:

$$F = \sum_j \frac{m_j}{m} \max_i F(i, j) \quad (8.15)$$

Където максимумът се взима по всички клъстери i от всички нива, m_j е броят на обекти в класа j, а m е общия брой обекти.

8.8. Оценка на значение на мерки за валидността на клъстери

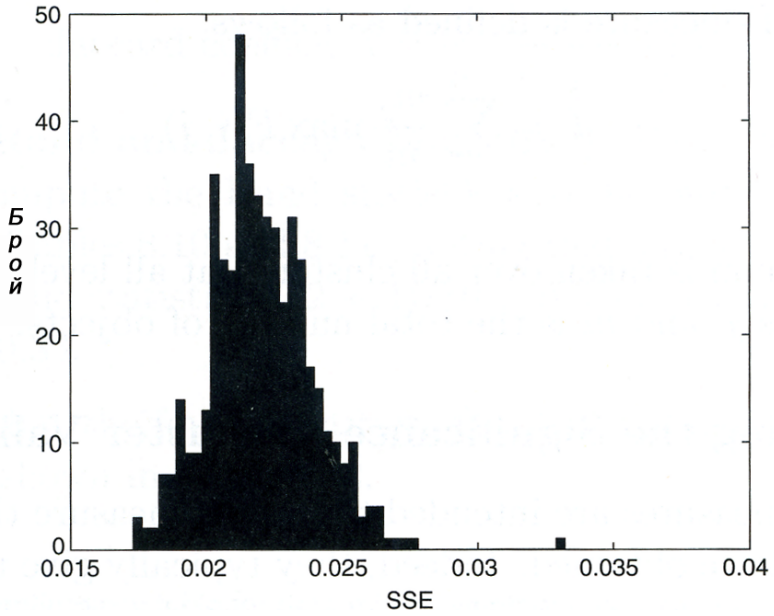
Мерките за оценяване на валидността на клъстери са предназначени да помогнат в преценката на качеството на получените при клъстеризацията клъстери. Обикновено, в качеството на такава оценка те ни дават само едно число. Интерпретацията на това число и на неговата значимост съвсем не е лека задача.

В повечето от случаите помощта за решаване на тази задача можем да получим от известните минимални и максимални стойности на избраната от нас мярка за оценяване на клъстери. Например, по определение, чистотата равна на 0 е лош показател, докато равна на 1 е добър, най-малко в тези случаи, когато вярваме на етикетите на класа и искаме получената клъстеризация да отговаря на структурата на класове. Също нулевите стойности на ентропията и средно-квадратичната грешка са също добри.

Обаче, понякога дори когато имаме максимални и минимални стойности на избраната мярка, нейните междинните стойности се нуждаят от интерпретация. Обикновено в тези случаи се използва статистическия подход. С други думи, ние опитваме да преценим, доколко вероятно е, че получената стойност може да бъде получена по случаен начин. Стойността е добра, ако тя е необикновена, т.е. ако е много малко вероятно тя да бъда получена случайно. Мотивацията за този подход е, че ние се интересуваме само от такива клъстери, които отразяват неслучайните структури в данни, а такива структури трябва да генерират необикновено високи (или низки) стойности на избраната от нас мярка за валидността на клъстери.

За илюстрацията на това, как този принцип действа, да разгледаме един пример, базиран на алгоритъма k-Means и мярката SSE. Да предположим, че искаме да измерим, доколко добре са отделени клъстери, показани на Фиг. 8-4 в сравнение със случайни

данни. За целта ще генерираме няколко случайни множества от по 100 точки, които имат същия диапазон, както и точки от трите клъстера. След това ще намерим по три клъстера във всяко от множество чрез прилагането на k-Means и да акумулираме получените стойности на SSE. Чрез използване на разпределение на тези стойности можем да приближим вероятността на стойността на SSE на оригиналните клъстери. На Фиг. 8-11 е показана хистограма на SSE за 500 случайни извадки. Най-ниската стойност на SSE е 0.0173. За клъстерите, показани на Фиг. 8-4 стойността на SSE е 0.0050. Следователно, можем да твърдим, че клъстеризацията, представена на Фиг. 8-4 може да се получи по случаен начин с вероятността по-малка от 1%.



Фиг. 8-11. Хистограма на SSE за 500 случайни множества от данни

И така, оценяването на качеството на клъстеризация (дали направлявано или не) е нещо повече от изчисляване на някакво число. За изключение на случаи, когато получената стойност има естествена интерпретация, базирана на дефиницията на използваната мярка, ние се нуждаем от интерпретацията на получената стойност. Ако нашата мярка е определена по такъв начин, че по-низките стойности указват на по-добри клъстери, то можем да използваме статистически подход, за да преценим, дали изчислената стойност е необикновено низка, използвайки разпределението на използваната мярка.

И на края, дори ако избраната мярка за качество се ползва като една относителна мярка, т.е. за сравнение на две клъстеризации, ние пак се нуждаем от оценяване на важността на изчислената разлика в оценките на двете клъстеризации. Макар, че винаги една стойност ще бъде по-добра от друга, обикновено е доста тежка задача да определим, доколко тази разлика е значителна. Обърнете внимание, че има два аспекта на тази значимост: 1) дали разликата е статистически значима (т.е. повторяема) и 2) дали размерът на разликата е от значение по отношение на конкретно приложение. Например, в повечето приложения разликата от 0.1% няма да се разглежда като значима, дори ако тя е напълно повторяема (т.е. значима от статистическа гледна точка).