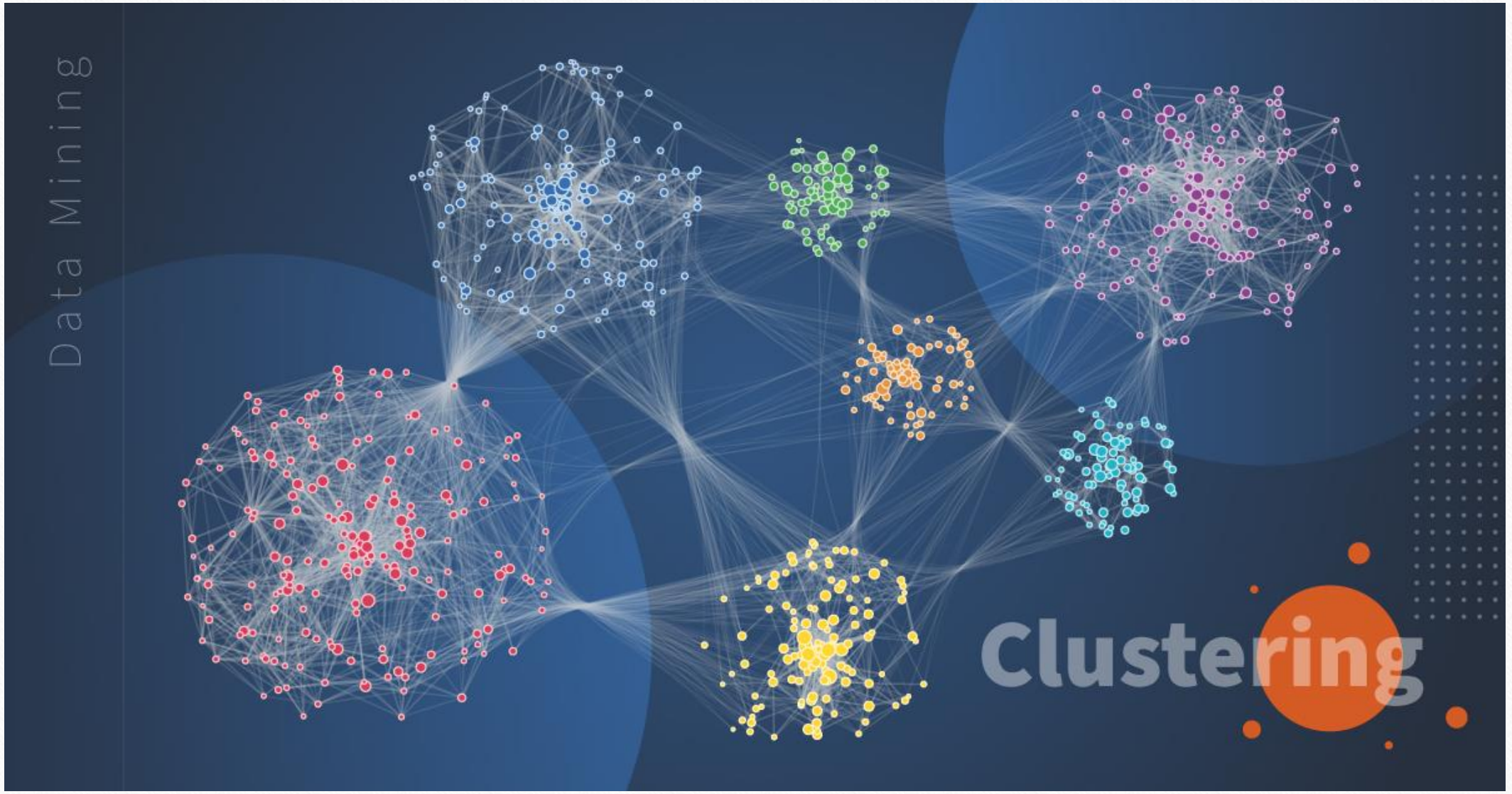
یادگیری بدون ناظر

تحلیل خوشه‌ای

یادگیری بدون ناظر:

- اگر در پایگاه داده ها، متغیر (یا فیلد) بر چسب دار وجود نداشته باشد برای مدلسازی این نوع از پایگاه داده ها فقط می توان از روش های یادگیری بدون ناظر استفاده کرد.

- از جمله این روش ها می توان به **تحلیل خوشه ای** و روش های پیشرفته آن اشاره کرد.


Data Mining
Clustering

آنالیز خوشه‌بندی

- آنالیز خوشه‌بندی یک روش قدیمی برای دسته‌بندی داده‌ها است که اولین بار توسط تریون در سال 1935 پیشنهاد شد.
- امروزه به دلیل استفاده از ابر کامپیوترها برای محاسبات سنگین، پیشرفت‌های زیاد و قابل توجهی برای آن صورت گرفته است.
- تفاوت روش خوشه‌بندی با دسته‌بندی در آن است که در این روش متغیر هدف و یا متغیربرچسب‌دار نداریم، لذا این روش تحلیل داده در دسته روش‌های بدون ناظر قرار می‌گیرد.
- به عبارت دیگر متغیرهای موجود به دو دسته مستقل و وابسته تقسیم نمی‌شوند.
- **هدف اصلی در خوشه بندی** تقسیم بندی اشیاء به گونه ای است که بیشترین شباهت در یک گروه و بیشترین تفاوت با اشیاء گروه های دیگر را دارا باشد.

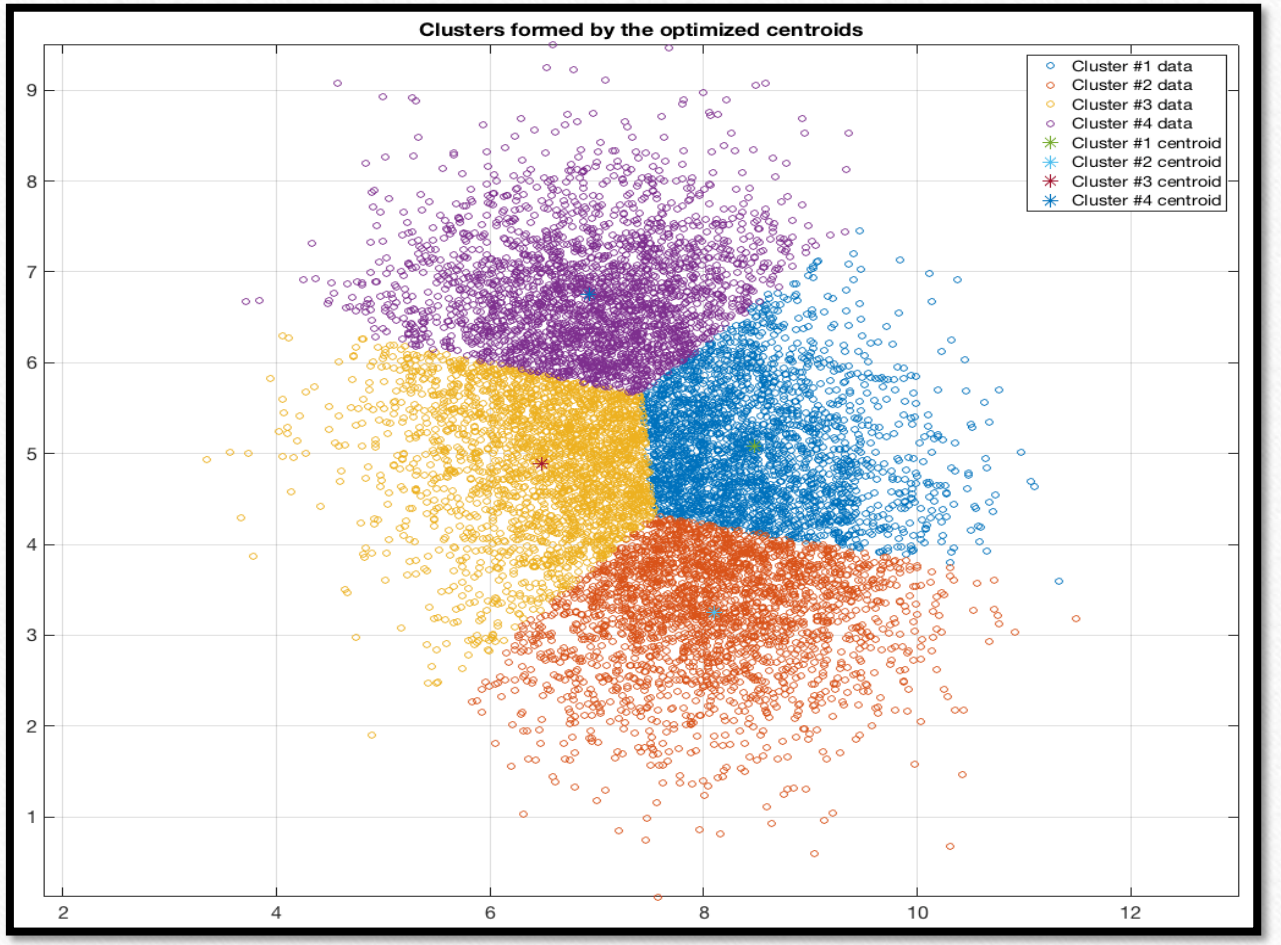
- تحلیل خوشه‌ای فرآیند تقسیم مجموعه‌ای از داده‌ها یا مشاهدات در زیر مجموعه‌ها می باشد، **هر زیر مجموعه یک خوشه است**. همه اعضای هر خوشه **مشابه** یکدیگرند و با اعضای دیگر خوشه‌ها **تفاوت** دارند. امکان این تقسیم‌بندی توسط انسان‌ها وجود ندارد و باید توسط الگوریتم‌های خوشه‌بندی صورت گیرد.
- از روش تحلیل خوشه ای برای شناخت گروه‌های ناشناخته در داده‌ها استفاده می‌شود. از روش تحلیل خوشه‌ای در زمینه‌های مختلف از جمله هوش تجاری، روش‌های جستجو اینترنتی، زیست شناسی، ... استفاده می شود.
- **مثال:** در یک شرکت مشاوره‌ای با خوشه بندی پروژه‌ها می توان آن دسته از پروژه‌هایی که از نظر ویژگی‌های تعریف شده برای پروژه‌ها، شباهت دارند دریک خوشه قرار گیرند.

- تحلیل خوشه‌ای یکی از روش‌های آماری است که برای کاهش داده‌ها و پیدا کردن کروه‌های مفید بسیار کاربرد دارد.
- توجه شود در خوشه‌بندی، داده‌ها در گروه‌هایی قرار می‌گیرند که از نظر محقق در آن زمینه مشابه باشند. اگر زاویه دید محقق تغییر کند، گروه‌ها در خوشه‌بندی تغییر می‌کنند.
- یکی از دلایل استفاده از خوشه‌بندی کشف ساختارهای جدید یا الگوهای جدید می‌باشد که به صورت طبیعی در داده‌ها وجود دارد اما بایستی توسط تحلیل خوشه‌ای کشف شود.

(Clustering) یا خوشه بندی ازجمله الگوریتم های یاد گیری مدل به حساب می‌آید. الگوریتم خوشه‌بندی رکوردهایی را که ویژگی‌های نزدیک به هم و مشابه دارند را در یک خوشه قرار می‌دهد. وقتی با یک مجموعه کوچک از صفات روبرو باشیم خوشه بندی به سادگی قابل اجرا است

- ✓ برای مثال در یک مجموعه از خودکارهای آبی، مشکی، قرمز و سبز به راحتی می‌توانیم آن‌ها را در 4 دسته قرار دهیم اما اگر در همین مجموعه ویژگی‌های دیگری مثل سایز، شرکت سازنده، وزن، قیمت و... مطرح باشد کار کمی پیچیده می‌شود. حال فرض کنید در یک مجموعه متشکل از هزاران رکورد وصدها ویژگی قصد دسته بندی دارید، چگونه باید این کار را انجام دهید؟

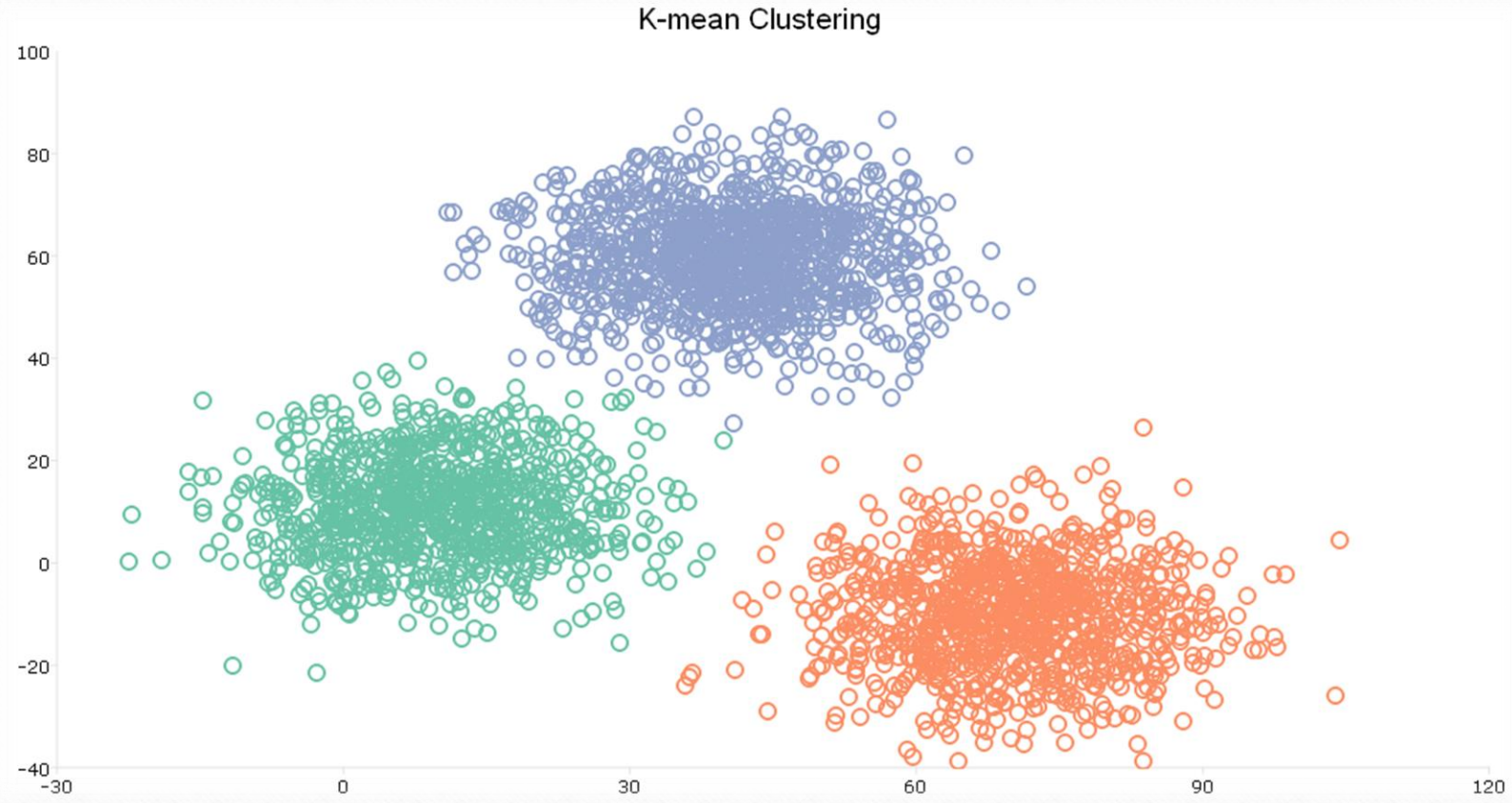
خوشه بندي در واقع تقسيم بندي يك جمعيت ناهمگون به تعدادي زير مجموعه همگون است.



- تقسیم داده‌ها به گروه‌ها یا خوشه‌های معنادار به طوری که محتویات هر خوشه ویژگی‌های مشابه و در عین حال نسبت به اشیاء دیگر در سایر خوشه‌ها غیر مشابه باشند را خوشه بندی می‌گویند. از این الگوریتم در مجموعه داده‌های بزرگ و در مواردی که تعداد ویژگی‌های داده زیاد باشد استفاده می‌شود.
- زمانی که با یک مجموعه کوچک از خصوصیات سروکار داریم، خوشه‌بندی یک عمل ساده‌ای است که می توانیم آن را انجام دهیم. اما زمانی که خصوصیات رشد می کنند مشکلات خوشه بندی افزایش پیدا کرده و حتی ممکن است از طریق ذهن آدمی غیر ممکن باشد. عمل خوشه بندی فقط در 5 یا 6 ویژگی برای افراد نظریه پرداز که فهم عمیقی از داده ها دارند امکان پذیر است. اما مجموعه داده‌های مدرن، عموماً شامل ده ها (اگر نگوییم صدها) ویژگی هستند.

- خوشه بندی، اشیاء را براساس ویژگی‌هایی که با هم دارند گروه‌بندی می‌کند.
- هدف اصلی در خوشه بندی تقسیم بندی اشیاء به گونه‌ای است که بیشترین شباهت در یک گروه و بیشترین تفاوت با اشیاء گروه‌های دیگر را دارا باشد.
- به عنوان تعریف ساده‌تر می‌توان گفت که اشیاء در خوشه مخصوص خود دارای بیشترین شباهت و در برابر اشیای متعلق به خوشه‌های دیگر دارای بیشترین تفاوت هستند.
- در روش خوشه‌بندی هیچ خوشه‌ای از ابتدا مشخص نیست. به عبارتی متغیرهای مورد نظر در ابتدا کاملاً از هم مستقل فرض می‌شوند.
- یکی از کاربردهای خوشه‌بندی **تشخیص نقاط پرت** بعد از تشکیل خوشه‌هاست.

در خوشه‌بندی، هدف تقسیم داده‌ها به گروه‌های مختلف است که با رنگ‌های مختلف در اینجا نشان داده شده‌اند.



- در این قسمت تفاوت تجزیه و تحلیل خوشه ای با دیگر تکنیک های گروه بندی را بیان می کنیم. در واقع گروه بندی اشیاء یا داده ها را به کلاس‌های نام گذاری شده تخصیص می‌دهد، در دسته بندی هر شیء دارای یک سرپرست و یا ناظر می‌باشد ولی می‌توان گفت خوشه بندی یک نوع گروه بندی بدون سرپرست یا ناظر است یعنی اشیاء بر اساس شباهت‌هایی که باهم دارند تقسیم می‌شوند و نه براساس معیارهای از پیش تعیین شده.
- به همین دلیل به خوشه بندی گاهی اوقات دسته بندی بدون ناظر نیز می‌گویند.
- در داده کاوی هنگامی که از اصطلاح دسته بندی استفاده می‌شود منظور همان دسته بندی به صورت با ناظر می‌باشد. در واقع استفاده از دسته بندی، استفاده از تکنیک‌های ساده ای برای گروه بندی داده‌ها می‌باشد.

الف) خوشه بندی با دو خوشه

ب) خوشه بندی با چهار خوشه

ج) خوشه بندی با شش خوشه

کاربرد های خوشه بندی:

- تشخیص داده های دور افتاده (مقادیری که با خوشه ها فاصله زیاد دارند).
- گروه بندی داده ها برای پیش بینی

خوشه بندی شاخه ای از علم آمار است، که یکی از موضوعات بسیار پرکاربرد در داده کاوی می باشد.

مراحل خوشه‌بندی

- انتخاب معیار مناسب برای تشخیص شباهت‌ها یا نزدیکی داده‌ها
- انتخاب روش تجزیه و تحلیل خوشه‌ها
- تصمیم‌گیری در مورد تعداد خوشه‌ها
- تفسیر نتایج خوشه‌ها و یا گروه‌ها

- **از جمله روش‌های خوشه‌بندی می‌توان به موارد زیر اشاره کرد:**

1. روش افرازبندی
2. روش سلسله مراتبی
3. روش‌های مبتنی بر تابع چگالی
4. روش‌های مبتنی بر مدل
5. روش Grid-based

روش افراز بندی

تحلیل خوشه‌ای

روش افرازبندی

- در این روش k افراز (k خوشه) مجزا درست می‌شود.
- هر افراز یک خوشه را درست میکند یعنی مشاهدات (n مشاهده) در این k خوشه قرار می‌گیرند. به طوریکه:

I. هر گروه یا هر خوشه حداقل یک عضو داشته باشد.

II. هر عضو(هر واحد) فقط به یک خوشه تعلق داشته باشد.

✓ به عبارت دیگر در روش افرازبندی، خوشه‌ها کاملاً از هم مستقل بوده و با یگدیگر هم‌پوشانی ندارند.

انواع الگوریتم‌های مربوط به روش افرازبندی

1) روش k-means
2) روش k-medoids
3) روش k-medians
4) روش Fuzzy C-means
5) روش K-SVD و ...

الگوریتم k-means

- الگوریتم k میانگین یکی از ساده‌ترین و محبوب‌ترین الگوریتم‌هایی است که در « داده کاوی » بخصوص در حوزه « مدل یادگیری بدون ناظر » کار می‌رود.
- معمولا در حالت چند متغیره، باید از ویژگیهای مختلف اشیا به منظور خوشه کردن آن‌ها استفاده کرد. به این ترتیب با داده‌های چند بعدی سروکار داریم که معمولاً به هر بعد از آن، ویژگی یا خصوصیت گفته می‌شود.
- با توجه به استفاده از **توابع فاصله** مختلف که در این جا مطرح می‌شود، ممکن است بعضی از ویژگی‌های اشیا **کمی** و بعضی دیگر **کیفی** باشند. به هر حال آنچه اهمیت دارد روشی برای اندازه گیری **میزان شباهت** یا **عدم شباهت** بین اشیاء است که باید در روش‌های خوشه بندی لحاظ شود.

- در این الگوریتم داده‌ها به تعداد خوشه‌هایی از پیش تعیین شده (k معلوم) تقسیم می‌شوند. به طوریکه برای هر خوشه، مرکز خوشه به صورت مقدار اولیه درنظر گرفته شده و سپس فاصله هر نقطه تا مرکز محاسبه می‌شود و مشاهداتی که به هر مرکز نزدیک‌تر باشند، در خوشه مربوطه قرار گرفته یا تخصیص داده می‌شوند.
- پس از تخصیص هر مشاهده به یکی از خوشه‌ها و تشکیل خوشه‌ها، برای هر خوشه یک نقطه جدید به عنوان مرکز خوشه محاسبه می‌شود.
- در این الگوریتم مرکز خوشه، میانگین مشاهدات موجود در هر خوشه است.
- این فرآیند تا جایی تکرار می‌شود که تغییر در میانگین خوشه‌ها منجر به تغییر در مشاهدات درون هر خوشه نشود.
- **توجه شود که اگر مراکز اولیه خوشه‌ها به درستی انتخاب نشود، خوشه‌های حاصل شده در پایان الگوریتم کیفیت مناسبی نخواهند داشت و این یکی از ضعف‌های این الگوریتم است.**

توابع فاصله:

- هدف خوشه بندی قرار گرفتن مشاهدات یا رکوردهای نزدیک به هم در یک خوشه باشد، به طوریکه بین مشاهدات درون خوشه کمترین فاصله و بین خوشه ها بیشترین فاصله باشد. خوشه بندی را مبتنی بر فاصله می گویند.

انواع توابع فاصله

1. **فاصله اقلیدوسی:**

- با استفاده از «**فاصله اقلیدسی**» **Euclidean Distance** کوتاه‌ترین فاصله بین دو نقطه براساس رابطه فیثاغورث، محاسبه می‌شود. اگر x و y دو نقطه با p مؤلفه باشند، فاصله اقلیدسی بین این دو به صورت زیر قابل محاسبه است:

$$D_{euc} = \left(\sum_{i=1}^{p} (x_i - y_i)^2 \right)^{\frac{1}{2}}$$

✓ فاصله اقلیدسی نامنفی است.

✓ وجود رابطه انعکاسی برای فاصله اقلیدسی اثبات می شود.

✓ رابطه مثلثی برای فاصله اقلیدسی برقرار است.

2. **فاصله منهتن:**

- اگر به جای مربع فاصله بین مولفه‌ها، از قدرمطلق فاصله بین مولفه‌های نقاط استفاده شود، تابع فاصله را **Manhatan** «منهتن» می‌نامند. این نام به علت تقاطع منظم خیابان‌ها درمحله منهتن نیویورک انتخاب شده است.
- اگر **x** و **y** نقطه با **p** مؤلفه (**p**بعدی) باشند، شیوه محاسبه فاصله منهتن به صورت زیرخواهد بود:

$$D_{man} = \sum_{i=1}^{p} |x_i - y_i|$$

✓ **نکته**: از آنجایی که میانه مولفه‌ای، نقطه ای است که فاصله منهتن را کمینه می کند، در بعضی از الگوریتم‌های خوشه بندی از میانه به عنوان مرکز هر خوشه استفاده می‌شود.

3. فاصله مینکوفسکی:

- **فاصله مینکوفسکی** حالت کلی تری از فاصله اقلدیسی و منهتن است. اگر **A** و **B** دو نقطه در فضای **p** بعدی باشند، فاصله مینکوفسکی برای آن‌ها به صورت زیر محاسبه می شود. پارامتر فاصله مینکوفسکی در اینجا **d** درنظر گرفته شده است، این فاصله به صورت زیر نشان داده می شود.

$$D_{mink}(A, B; d) = \left(\sum_{i=1}^{p} |x_i - y_i|^d \right)^{\frac{1}{d}}$$

مثال از الگوریتم k-means

- مشاهدات زیر مفروض هستند.
- {2 , 4, 10, 12, 3, 20, 30, 11, 25}

مطلوبست الگوریتم k-means برای k=2 و مقادیر اولیه $m_{01} = 3$ و $m_{02} = 4$.

	خوشه اول	خوشه دوم
مرحله اول	$m_{01} = 3$ $\{2,3\}$	$m_{02} = 4$ $\{4,10,12,3,20,30,11,25\}$
مرحله دوم	$m_{11} = 2.5$ $\{2,3,4\}$	$m_{12} = 16$ $\{10,12,20,30,11,25\}$
مرحله سوم	$m_{21} = 3$ $\{2,3,4,10\}$	$m_{22} = 18$ $\{12,20,30,11,25\}$
مرحله چهارم	$m_{31} = 4.75$ $\{2,3,4,10,11,12\}$	$m_{32} = 19.6$ $\{20,30,25\}$
مرحله پنجم	$m_{41} = 7$ $\{2,3,4,10,11,12\}$	$m_{42} = 25$ $\{20,30,25\}$

نقاط ضعف الگوریتم k-means

1. این روش زمانی کاربرد دارد که بتوان مرکز خوشه را حدس زد. چرا که جواب نهایی به انتخاب مراکز اولیه خوشه وابسته است.
2. روال خاصی برای تعیین مقدار k و مراکز اولیه خوشه وجود ندارد.
3. این روش برای کشف خوشه‌های پیچیده مناسب نیست.
4. این روش نسبت به داده‌های پرت(داده‌های دور از مرکز) حساس است چرا که داده‌های پرت به راحتی می‌توانند مرکز خوشه را تغییر دهند و نتایج را نامناسب نشان دهند.
5. امکان تولید خوشه‌های خالی توسط الگوریتم وجود دارد.
6. تفسیر نتایج حاصل از خوشه‌بندی نیاز به تجربه و افراد خبره در این زمینه دارد.
7. انتخاب روش‌های صحیح برای تعیین فواصل رکوردها تا مراکز خوشه کار آسانی نیست.

راه حل‌های پیشنهادی برای رفع ضعف‌های الگوریتم k-means

- راه حل‌های رفع **مشکل انتخاب مراکز اولیه** در الگوریتم k-means:

1. اجرای چندین باره این الگوریتم به ازای مقادیر اولیه متفاوت می‌باشد.
2. راه حل دیگر انتخاب تعداد مراکز اولیه بیش از تعداد خوشه‌ها و سپس ادغام کردن آن‌ها باتوجه به مقدار k است.
3. سومین راه حل انتخاب مقادیری است که بیشترین جداکنندگی را داشته باشند.

- برای رفع **ضعف الگوریتم k-means در تولید خوشه‌های خالی** راه حل‌های زیر می‌توانند کارآمد باشند:

1. مشاهده‌ای که دارای دورترین فاصله از مرکز خوشه است را به عنوان یک خوشه جدید انتخاب کنیم و خوشه‌های خالی را پر کنیم.
2. در صورتیکه مرکز خوشه خالی یکی از رکوردهای ما باشد، می‌توان آن را به عنوان داده پرت تشخیص داد و از میان مشاهدات خارج کرد.
3. اگر چندین خوشه خالی وجود داشته باشد، می‌توان با تکرار روش‌های فوق ضعف خوشه‌های خالی را برطرف کرد.

✓ **اگر درمیان داده‌ها نقاط پرت زیادی وجود داشته باشد در این صورت پیشنهاد می‌شود به جای استفاده از میانگین داده‌ها برای تعیین مرکز خوشه، از میانه خوشه‌ها استفاده شود که به این الگوریتم، الگوریتم k-median گفته می‌شود.**

بهبود کیفی خوشه‌های تولیدشده:

1. **شکستن خوشه‌ها به خوشه‌های کوچکتر:**

در صورتی که رکوردهایی وجود داشته باشد که فواصل آن‌ها از تا مرکز از سایر مشاهدات داخل خوشه بیشتر باشد، توصیه به شکستن آن خوشه می‌شود.

2. **ادغام خوشه‌های مشابه یا نزدیک:**

خوشه‌هایی که مجموع فواصلشان تا مرکز آن‌ها از سایر خوشه‌ها کمتر باشد(برای مثال دو الی چند خوشه به هم بسیار نزدیک باشند)، توصیه به ادغام این خوشه‌ها می‌شود.

3. **حذف خوشه‌های کوچک:**

در صورتی که بعضی از نقاط پرت باشند، خوشه‌های کوچک ایجاد شده توسط این نقاط ممکن است نامناسب بوده و توصیه به حذف خوشه می‌گردد.

خوشه بندی سلسله مراتبی

تحلیل خوشه‌ای

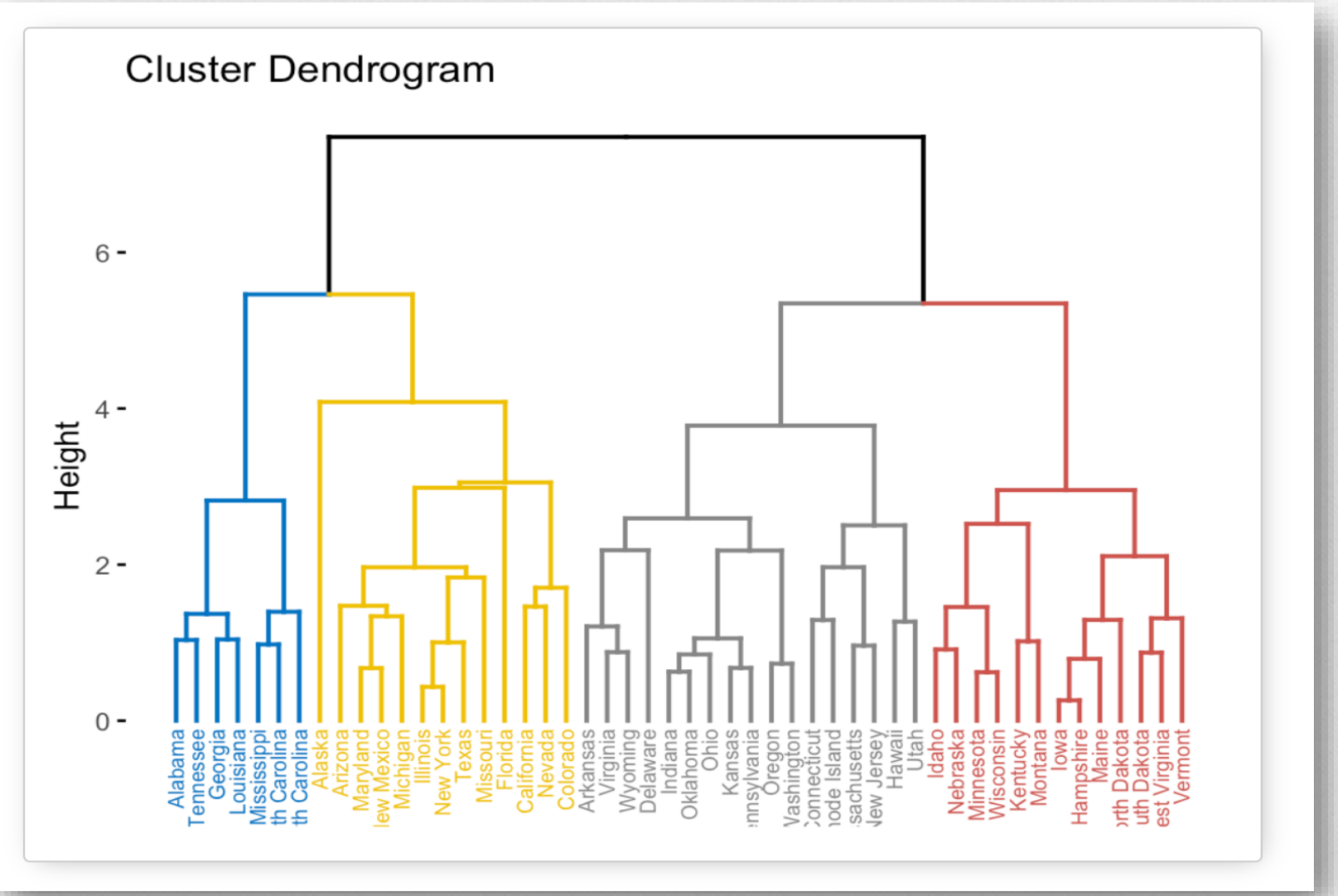
خوشه بندی سلسله مراتبی

- با توجه به آن که روش‌های خوشه‌بندی مطرح شده تا حد زیادی پاسخ‌گوی نیاز ما هستند، اما در بعضی از موارد هدف تقسیم‌بندی داده‌ها در گروه‌هایی است که سطوح متفاوتی داشته و نسبت به یکدیگر در سلسله مراتب مختلفی قرار می‌گیرند.
- از این رو روش‌های خوشه‌بندی سلسله مراتبی از اهمیت بیشتری برخوردار هستند.
- مثال) مدیر منابع انسانی شرکت الکترونیکی، کارکنان شرکت خود را در گروه‌هایی مانند مدیران اجرایی، مدیران و کادر اداری سازماندهی می‌کند. کادر اداری را نیز به زیرگروه‌هایی مانند کارمندان ارشد، کارمندان و کارآموزان تقسیم می‌کند. با این تقسیم‌بندی کسب اطلاعات و خلاصه‌سازی و استخراج اطلاعات راحت‌تر می‌باشد.

- برعکس خوشه بندی تفکیکی که اشیاء را در گروه های مجزا تقسیم می کند، «خوشه بندی سلسله مراتبی» در هر سطح از فاصله، نتیجه خوشه‌بندی را نشان می دهد. این سطوح به صورت سلسله مراتبی هستند.

- برای نمایش نتایج خوشه بندی به صورت سلسله مراتبی از **« درختواره »** **Dendrogram** استفاده می شود. این شیوه، روشی موثر برای نمایش نتایج خوشه بندی سلسله مراتبی است.

- روش‌های خوشه بندی سلسله مراتبی خود نیز شامل چندین روش می‌باشند که از جمله آن‌ها می‌توان به خوشه‌بندی سلسله مراتبی چندمرحله‌ای، خوشه بندی سلسله مراتبی با استفاده از مدل‌سازی پویا، خوشه‌بندی سلسله مراتبی احتمالی و ... اشاره کرد.


Cluster Dendrogram
Height
6
4
2
0
Alabama
Tennessee
Georgia
Louisiana
Mississippi
th Carolina
th Carolina
Alaska
Arizona
Maryland
lew Mexico
Michigan
Illinois
New York
Texas
Missouri
Florida
California
Nevada
Colorado
Arkansas
Virginia
Wyoming
Delaware
Indiana
Oklahoma
Ohio
Kansas
nnsylvania
Oregon
Vashington
Connecticut
ıode Island
sachusetts
lew Jersey
Hawaii
Utah
Idaho
Nebraska
Minnesota
Wisconsin
Kentucky
Montana
Iowa
Hampshire
Maine
orth Dakota
uth Dakota
est Virginia
Vermont

بر خلاف تکنیک ها ی خوشه بندی که بر پایه افراز که تعداد خوشه ها به عنوان پارامتر ورودی توسط کاربر مشخص می شد ، در تکنیک های خوشه بندی سلسله مراتبی مجموعه داده ها و معیاری جهت ارزیابی تشابه به عنوان ورودی معین می شوند .

بسته به اینکه تحلیل این ساختار سلسله مراتبی از پایین به بالا (Bottom-up) و یا برعکس از بالا به پایین (Top-down) انجام شود ، عملیات اصلی این الگوریتم ها را می توان در دو دسته ی ادغام و تقسیم قرار داد . فرآیند ادغام و یا تقسیم چند خوشه در این تکنیک ها نقش مهمی را ایفا می کند . چرا که در صورت اخذ تصمیمی ضعیف جهت ادغام و تقسیم خوشه ها تکنیک توانایی برگشت و اصلاح آن را ندارد و این عمل باعث کاهش کیفیت در خوشه های تولید شده نهایی خواهد داشت .

الگوریتم های خوشه بندی سلسله مراتبی که یک فرآیند پایین به بالا را طی می کنند ، کارشان با قرار دادن هر نمونه داده در یک خوشه ی مجزا شروع می شود با ادغام خوشه ها الگوریتم تا جایی پیش می رود که یا کلیه ی نمونه ها در یک خوشه قرار گیرند و یا شرط از پیش تعیین شده ای به عنوان پایان اجرا مشخص شده باشد . اغلب تکنیک های موجود سلسله مراتبی متعلق به این دسته هستند .

در طرف دیگر الگوریتم های سلسله مراتبی قرار دارند که یک استرانژی بالا به پایین را دنبال می کنند . در این روش ها بر عکس تکنیک های قبلی در ابتدا کلیه نمونه ها در یک خوشه قرار می گیرند . پس از آن با کمک از یک معیار تشابه در چند مرحله بصورت سلسله مراتبی این خوشه به خوشه های کوچکتر تقسیم می شود . در این الگوریتم ها نیز می توان کار را تا جایی ادامه داد تا هر نمونه در یک خوشه قرار بگیرد و یا اینکه شرطی را جهت پایان اجرای الگوریتم معین نمود . در هر دو دسته از الگوریتم های سلسله مراتبی کاربر می تواند تعداد خوشه های تولید شده نهایی را بعنوان یک شرط پایانی مشخص کنند .

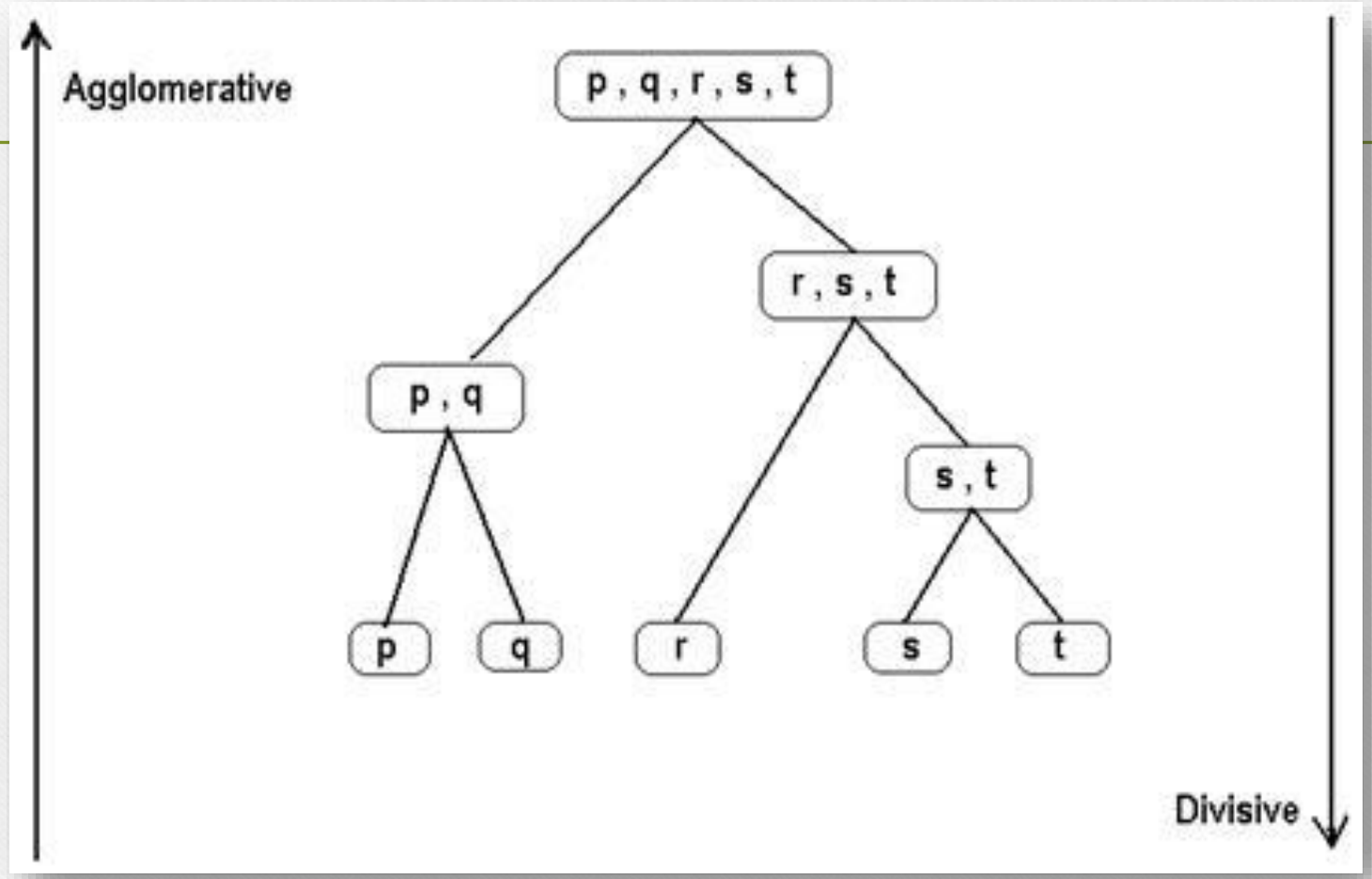
1.بالا به پايين (Top-Down) يا تقسيم كننده(Divisive)

در اين روش ابتدا تمام داده‌ها به عنوان يک خوشه در نظر گرفته مي‌شوند و سپس در طي يک فرايند تكراري در هر مرحله داده‌هايي شباهت كمتري به هم دارند به خوشه‌هاي مجزايي شكسته مي‌شوند و اين روال تا رسيدن به خوشه‌هايي كه داراي يک عضو هستند ادامه پيدا مي‌كند.

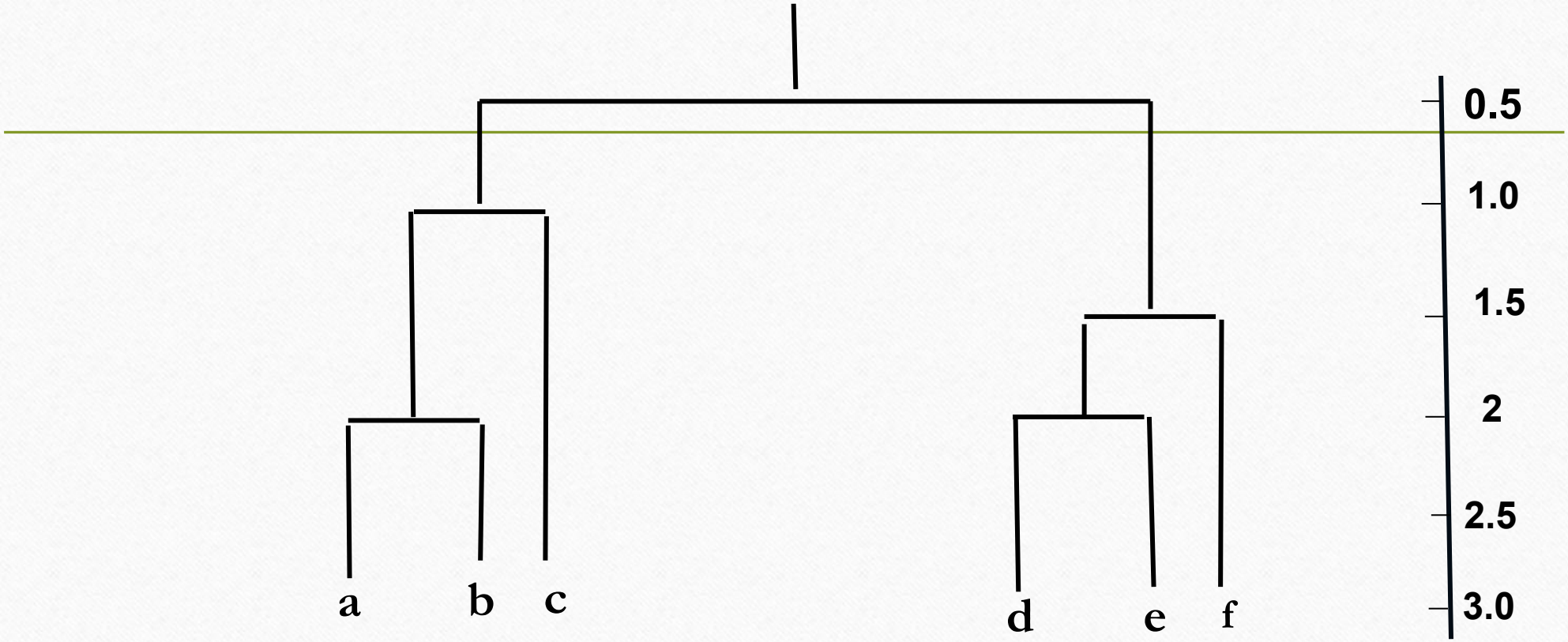
2.پايين به بالا (Bottom-Up) يا متراكم شونده(Agglomerative)

در اين روش ابتدا هر داده‌ها به عنوان خوشه‌اي مجزا در نظر گرفته مي‌شود و در طي فرايندي تكراري در هر مرحله خوشه‌هايي كه شباهت بيشتري با يكديگر با يكديگر تركيب مي‌شوند تا در نهايت يک خوشه و يا تعداد مشخصي خوشه حاصل شود. از انواع الگوريتمهاي خوشه‌بندي سلسله مراتبي متراكم شونده رايج مي‌توان از الگوريتمهايSingle-Link ، Average-Linkو Complete-Link نام برد. تفاوت اصلي در بين تمام اين روشها به نحوة محاسبة شباهت بين خوشه‌ها مربوط مي‌شود.

تفاوت بين روشهاي بالا به پايين با روشهاي پايين به بالا

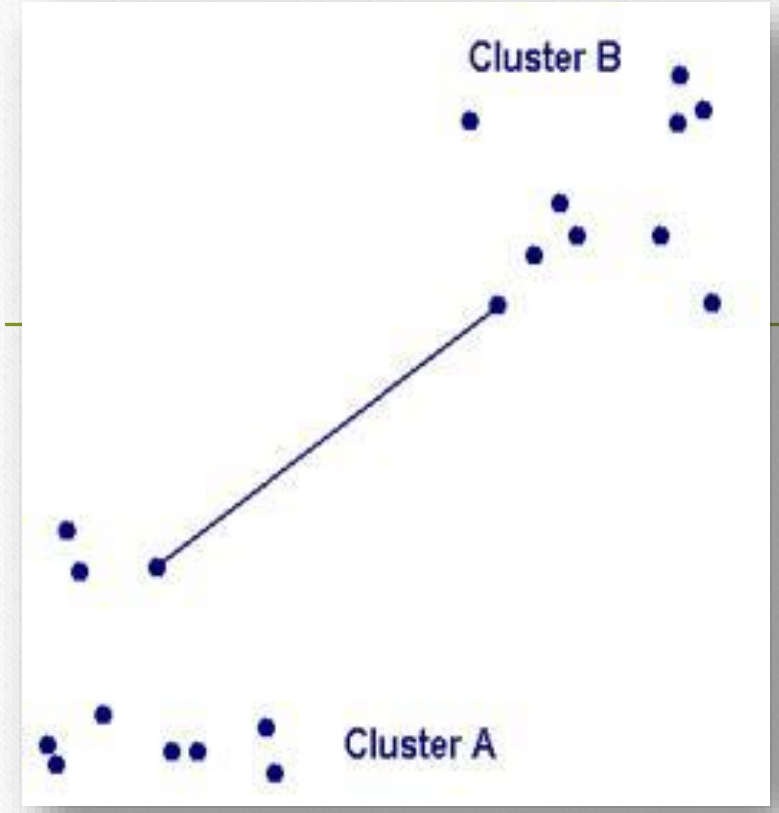


معمولا فرآیند خوشه ای سلسله مراتبی توسط یک نمودار با نام دندوگرام نمایش داده می شود که مثالی از آن در شکل زیر ملاحظه می شود :



این نمودار ادغام و تقسیم خوشه ها در هر مرحله نمایش می دهد . محور عمودی کنار نمودار، مقادیر مقیاس میان خوشه ها را نشان می دهد . برای مثال همان طور که مشاهده می شود هنگامی تشابه میان خوشه های $\{a,b,c\}$و $\{e,f,d\}$ تقریبا برابر با 0.5 این دو خوشه با هم ادغام می شوند.

خوشه‌بندي با روش Single-Link



شباهت بين دو خوشه در روش Single-Link برابر است با كمترين فاصلة بين داده‌هاي دو خوشه

اين روش يكي از قديمي‌ترين و ساده‌ترين روشهاي خوشه‌بندي است و جزء روشهاي خوشه‌بندي سلسله مراتبي و انحصاري محسوب مي‌شود. به اين روش خوشه‌بندي، تكنيک نزديكترين همسايه (Nearest Neighbour) نيز گفته مي‌شود .در اين روش براي محاسبة شباهت بين دو خوشة A و B از معيار زير استفاده مي‌شود:

$$d_{AB} = min d_{ij} \qquad i \epsilon A, j \epsilon B$$

كه i يک نمونه داده متعلق به خوشة A و jيک نمونه دادة متعلق به خوشة B مي‌باشد.

در واقع در اين روش شباهت بين دو خوشه، كمترين فاصلة بين يک عضو از يكي با يک عضو از ديگري است. در شكل روبرو اين مفهوم بهتر نشان داده شده است.

مثال: فرض می شود که 5 نمونه داده موجود است که هر یک از آنها دارای دو ویژگی X و Y می باشد

	O_1	O_2	O_3	O_4	O_5
X	1	5	6	1	4
Y	2	2	1	1	1

	O_1	O_2	O_3	O_4	O_5
O_1	0				
O_2	4	0			
O_3	5.1	1.4	0		
O_4	1	4.1	5.0	0	
O_5	3.1	1.4	2.0	3.0	0

معیار ارزیابی فاصله اقلیدسی است .در میان مقادیر کوچکترین عدد یک است که نشان دهنده ی تشابه بالای O_1 و O_4می باشد بنابراین در این مرحله این دو نمونه می توانند در یک خوشه قرار گرفته و بعبارتی ادغام شوند .

Distance($\{O_1,O_4\}$,O_2)=min(d(O_1, O_2), d(O_4 ,O_2))=min(4.0,4.1)=4.0

Distance($\{O_1,O_4\}$,O_3)=min(d(O_1 , O_3),
d(O_4 ,O_3))=min(5.1,5.0)=5.0

Distance($\{O_1,O_4\}$,O_5)=min(d(O_1 , O_5),
d(O_4 ,O_5))=min(3.1,3.0)=3.0

	$\{O_1,O_4\}$	O_2	O_3	O_5
$\{O_1,O_4\}$	0			
O_2	4	0		
O_3	5.0	1.4	0	
O_5	**3.0**	1.4	2.0	0

در این مرحله دو مقدار **1.4** در ماتریس نشان دهنده ی نزدیکی نمونه های O_3 و O_5 به نمونه ی O_2 است می توان این سه نمونه را در یک خوشه ادغام کرد و کار را ادامه داد. اما اگر شرط الگوریتم این است که در هر مرحله فقط یکی از نمونه ها با نمونه یا خوشه ی دیگر ادغام شود ، مجبور خواهیم بود تا نمونه ی O_2 را با یکی نمونه های O_3 یا O_5 ادغام کنیم .

Distance($\{O_1,O_4\}$,$\{O_2,O_3\}$)=min(d(O_1, O_2), d(O_4 ,O_2), d(O_4 ,O_3), d(O_1, O_3))=4.0

Distance($\{O_1,O_4\}$,O_5)=min(d(O_1,O_5), d(O_4 ,O_5))=3.0

Distance($\{O_2,O_3\}$,O_5)=min(d(O_2,O_5), d(O_3 ,O_5))=1.4

نمونه ی O_2 را با O_3 ادغام و پس از آن ماتریس تشابه را بروز رسانی می کنیم

.

	$\{O_1, O_4\}$	$\{O_2, O_3\}$	O_5
$\{O_1, O_4\}$	0		
$\{O_2, O_3\}$	4.0	0	
O_5	3.0	1.4	0

ماتریس تشابه پس از اجرای مرحله دوم

با توجه به محتوای ماتریس تشابه در این مرحله نمونه ی O_5 با خوشه ی $\{O_2,O_3\}$ ادغام خواهد شد و ماتریس تشابه پس از این ادغام اصلاح می شود. اما فاصله همان 1.4 را نشان می دهد .

با توجه محتوای ماتریس تشابه در این مرحله نمونه ی O_5 با خوشه ی $\{O_2, O_3\}$ ادغام خواهد شد و ماتریس تشابه پس از این ادغام اصلاح می شود. اما فاصله همان 1.4را نشان می دهد.

	$\{O_1, O_4\}$	$\{O_2, O_3, O_5\}$
$\{O_1, O_4\}$	0	
$\{O_2, O_3, O_5\}$	**3.0**	0

ماتریس تشابه پس از اجرای مرحله سوم

بالاخره در مرحله نهایی دو خوشه در یک خوشه ادغام می شوند. در واقع کلیه نمونه ها در یک نمونه قرار می گیرند.

مثال: در این مثال 6 نمونه داده و ماتریس فاصلة بین آنها در جدول نشان‌داده شده است.

جدول 1: ماتریس فاصلة بین 6 نمونة داده

6	5	4	3	2	1	
8	12	24	13	4	0	1
10	11	22	10	0		2
9	3	7	0			3
18	6	0				4
8.5	0					5
0						6

در ابتدا هر داده به عنوان یک خوشه در نظر گرفته مي‌شود و یافتن نزدیکترین خوشه در واقع یافتن کمترین فاصلة بین داده‌هاي بالا خواهد بود. با توجه به جدول 1مشخص است که داده‌هاي 3 و 5 کمترین فاصله را دارا هستند. و در نتیجه آنها را با هم ترکیب کرده و خوشة جدیدي حاصل مي‌شود که فاصلة آن از سایر خوشه‌ها برابر است با کمترین فاصلة بین 3 و یا 5 از سایر خوشه‌ها .نتیجه در جدول 2نشان داده شده است .

با توجه به جدول 2مشخص است كه داده‌هاي 1 و 2 كمترين فاصله را دارا هستند. و در نتيجه آنها را با هم تركيب كرده و خوشة جديدي حاصل مي‌شود كه فاصلة آن از ساير خوشه‌ها برابر است با كمترين فاصلة بين 1 و يا 2 از ساير خوشه‌ها.

جدول 2: ماتريس فاصلة بين 5 خوشة حاصل از تكرار اول

	1	2	(3 و 5)	4	6
1	0	4	12	24	8
2		0	10	22	10
(3 و 5)			0	6	8.5
4				0	18
6					0

با توجه به جدول 3مشخص است كه خوشه‌هاي (3 و 5) و 4 كمترين فاصله را دارا هستند. و در نتيجه آنها را با هم تركيب كرده و خوشة جديدي حاصل مي‌شود كه فاصلة آن از ساير خوشه‌ها برابر است با كمترين فاصلة بين (3 و 5) و يا 4 از ساير خوشه‌ها.

جدول 3: ماتريس فاصلة بين 4 خوشة حاصل از تكرار دوم

	(1 و 2)	(3 و 5)	4	6
(1 و 2)	0	10	22	8
(3 و 5)		0	6	8.5
4			0	18
6				0

با توجه به جدول 4مشخص است که خوشه‌هاي (1 و 2) و 6 کمترين فاصله را دارا هستند. و در نتيجه آنها را با هم ترکيب کرده و خوشة جديدي حاصل مي‌شود که فاصلة آن از ساير خوشه‌ها برابر است با کمترين فاصلة بين (1 و 2) و يا 6 از ساير خوشه‌ها.

جدول 4: ماتريس فاصلة بين 3 خوشة حاصل از تکرار سوم

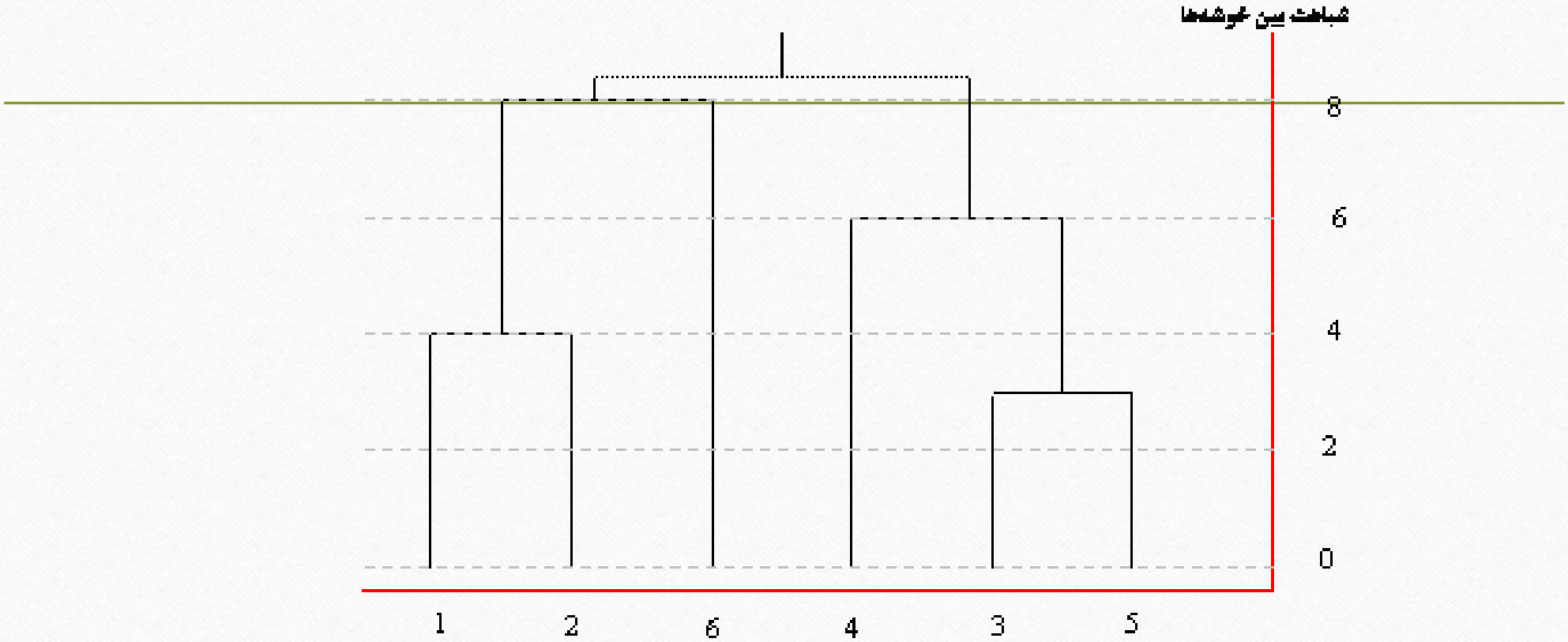
	(1 و 2)	(3، 4 و 5)	6
(1 و 2)	0	10	8
(3، 4 و 5)		0	8.5
6			0

جدول 5: ماتريس فاصلة بين 2 خوشة حاصل از تکرار چهارم

	(1، 2 و 6)	(3، 4 و 5)
(1، 2 و 6)	0	8.5
(3، 4 و 5)		0

در نهايت اين دو خوشة حاصل با هم ترکيب مي‌شوند.

نتیجه ترکیب نهایی در دندوگرام زیر نشان داده شده است .

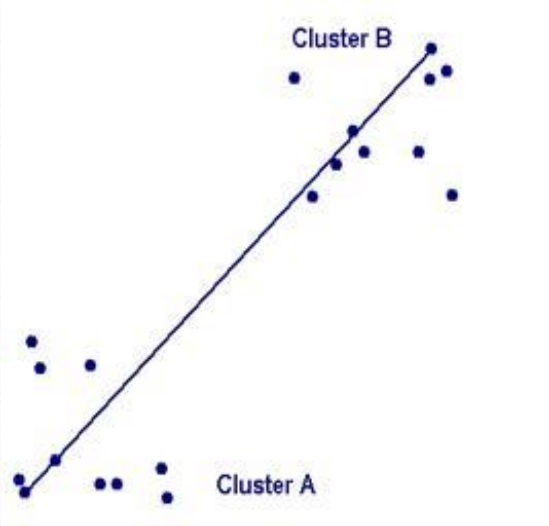


خوشه‌بندي با روش Complete-Link

اين روش همانند Single-Link جزء روشهاي خوشه‌بندي سلسله مراتبي و انحصاري محسوب مي‌شود. به اين روش خوشه‌بندي، تكنيک دورترين همسايه (furthest Neighbour) نيز گفته مي‌شود .در اين روش براي محاسبة شباهت بين دو خوشة A و B از معيار زير استفاده مي‌شود:

$$d_{AB} = max d_{ij} \qquad i \epsilon A, j \epsilon B$$

كه i يک نمونه داده متعلق به خوشة A و j يک نمونه دادة متعلق به خوشة B مي‌باشد. در واقع در اين روش شباهت بين دو خوشه بيشترين فاصلة بين يک عضو از يكي با يک عضو از ديگري است.



شباهت بين دو خوشه در روش Complete-Linkبرابر است با بيشترين فاصلة بين داده‌هاي دو خوشه.

مثال: در این قسمت سعي شده است تا در مثالي با فرض داشتن 5 نمونه داده که هر یک از آنها دارای دو ویژگی X و Y می باشد

	O_1	O_2	O_3	O_4	O_5
X	1	5	6	1	4
Y	2	2	1	1	1

	O_1	O_2	O_3	O_4	O_5
O_1	0				
O_2	4	0			
O_3	5.1	1.4	0		
O_4	1	4.1	5.0	0	
O_5	3.1	1.4	2.0	3.0	0

معیار ارزیابی فاصله اقلیدسی است .در میان مقادیر کوچکترین عدد یک است که نشان دهنده ی تشابه بالای O_1 و O_4می باشد بنابراین در این مرحله این دو نمونه می توانند در یک خوشه قرار گرفته و بعبارتی ادغام شوند .

Distance($\{O_1,O_4\}$,O_2)=max(d(O_1, O_2), d(O_4 ,O_2))=max(4.0,4.1)=4.1

Distance($\{O_1,O_4\}$,O_3)=max(d(O_1, O_3), d(O_4 ,O_3))=max(5.1,5.0)=5.1

Distance($\{O_1,O_4\}$,O_5)=max(d(O_1, O_5), d(O_4 ,O_5))=max(3.1,3.0)=3.1

	$\{O_1,O_4\}$	O_2	O_3	O_5
$\{O_1,O_4\}$	0			
O_2	**4.1**	0		
O_3	**5.1**	1.4	0	
O_5	**3.1**	1.4	2.0	0

در ماتریس فوق کوچکترین عدد **1.4** است که مربوط به تشابه میان نمونه O_2 با نمونه های O_3 و O_5 است . همانند مثال قبل در این لحظه بدلیل تشابه یکسان با دو انتخاب روبرو هستیم بدلیل اینکه نشان دهیم الگوریتم می تواند در هر مرحله چندین نمونه را در یک خوشه واحد قرار دهد، سه نمونه را در این مرحله در یک خوشه ادغام می کنیم. در مرحله پایانی همانند الگوریتم قبل اعضا دو خوشه ی باقیمانده در یک خوشه ادغام می شوند.

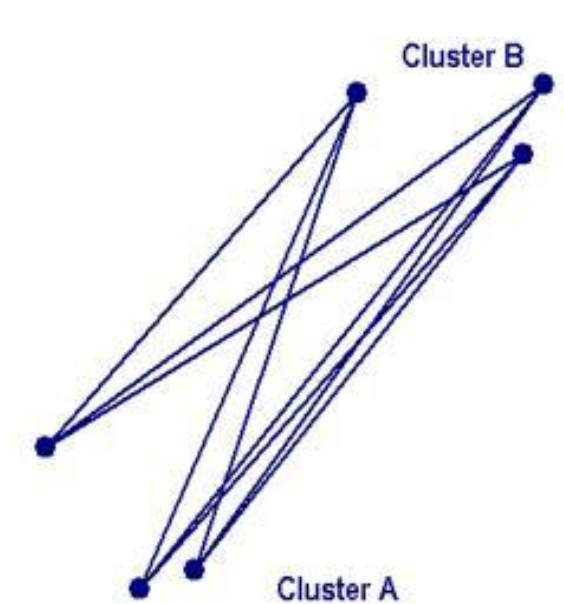
	$\{O_1,O_4\}$	$\{O_2,O_3,O_5\}$
$\{O_1,O_4\}$	0	
$\{O_2,O_3,O_5\}$	**5.1**	0

ماتریس تشابه نهایی

خوشه‌بندي با روش Average-Link

$$d_{AB} = \frac{\sum_{i \in A, j \in B} d_{ij}}{N_A N_B}$$

اين روش همانند Single-Link جزء روشهاي خوشه‌بندي سلسله مراتبي و انحصاري محسوب مي‌شود .از آنجا كه هر دو روش خوشه‌بندي Single-link و Complete-link بشدت به نويز حساس مي‌باشد، اين روش كه محاسبات بيشتري دارد، پيشنهاد شد. در اين روش براي محاسبة شباهت بين دو خوشة A و B از معيار زير استفاده مي‌شود:



كه i يک نمونه داده متعلق به خوشة A و j يک نمونه دادة متعلق به خوشة B مي‌باشد. و N_A تعداد اعضاء خوشة A و N_B تعداد اعضاء خوشة Bاست. در واقع در اين روش، شباهت بين دو خوشه ميانگين فاصلة بين تمام اعضاء يكي با تمام اعضاء ديگري است.

در روش میانگین، باید میانگین کلیه فواصل زوج نمونه ها را طوری که هر نمونه متعلق به یک خوشه است را محاسبه کرد.

$$\textbf{Distance}(C_i, C_j) = \frac{1}{|C_i| \times |C_j|} \sum_{o_i \in C_i, o_j \in C_j} \text{Distance}(o_i, o_j)$$

منظور از $|C_i|$ تعداد نمونه هایی است که در خوشه C_i قرار دارند .

	O_1	O_2	O_3	O_5
O_1	0			
O_2	**1.0**	0		
O_3	**11.0**	2.0	0	
O_5	**5.0**	**3.0**	**4.0**	0

با توجه به محتوای ماتریس، نمونه های O_1 و O_2 در مرحله ی اول ادغام می شوند . پس از آن چنانچه ماتریس تشابه تشکیل شود مشاهده می شود که نمونه های O_3 و O_4 می توانند در مرحله ی بعدی ادغام شوند . در نهایت دو خوشه ی $\{O_1 \text{ و } O_2\}$ و $\{O_3 \text{ و } O_4\}$ در یک خوشه قرار می گیرند .

نحوه ی محاسبه ی فاصله میان این دو خوشه پایانی به صورت زیر است:

$$\text{Distance}\left(\{O_1 \text{ و } O_2\}, \{O_3 \text{ و } O_4\}\right) = \frac{1}{4} \times \left(d(O_1 \text{ و } O_3) + d(O_1 \text{ و } O_4) + d(O_2 \text{ و } O_3) + d(O_2 \text{ و } O_4)\right) = \frac{1}{4}(11+5+2+3) = 5.25$$

روش های پیشرفته خوشه بندی

تحلیل خوشه‌ای

- در روش‌های خوشه‌بندی که تاکنون مطرح شده‌اند، هر داده می‌توانست تنها به یک خوشه اختصاص یابد. چنین روشی برای تخصیص داده‌ها به خوشه‌ها برای برخی از کاربردهای خوشه‌بندی مانند بازاریابی که هدف تخصیص هر مشتری به یک سیستم بازاریابی می‌باشد، مفید است.

✓ اما در برخی از کاربردهای خوشه بندی این امر چندان خوشایند نیست. لذا از روش‌های منطبق با ایده **فازی** برای خوشه‌بندی و تخصیص داده‌ها استفاده می‌شود.

✓ از جمله این روش‌ها می‌توان **خوشه‌بندی مبتنی بر مدل احتمالی** و **خوشه بندی فازی** و ... را نام برد.

- **مثال)** فرض کنید شرکت الکترونیکی فروشگاه آنلاینی دارد که علاوه بر خرید و فروش آنلاین، مشتریان می‌توانند نظرات خود را درباره محصولات در آنجا قرار دهند. این نظرات می‌توانند شامل چند محصول باشند و ممکن است محصولش حتی یک نظر هم نداشته باشد.

✓ هدف اصلی بررسی کننده نظرها، خوشه بندی نظرهای مختلف است. در این صورت تخصیص هر نظر به یکی از خوشه‌ها هدف این شرکت نمی‌باشد.

✓ در این صورت نیاز به روش خوشه‌بندی احساس می‌شود که نظرها شامل بیش از یک موضوع باشند. لذا آن‌ها را به بیش از یک خوشه اختصاص می‌دهد.