روش‌های استنباطی

تحلیل روش‌های پیشرفته‌آماری چند متغیره

۱- تحلیل خوشه‌ای

تحلیل خوشه‌ای مجموعه کثیری از داده‌ها را بر حسب فاصله آنها به خوشه یا دسته‌های کوچک‌تری تقسیم می‌کند. به‌این‌ترتیب که متغیر های که از همدیگر فاصله کمتری دارند را دریک گروه قرار می‌دهد. ازاین‌رو می‌توان گفت هدف اصلی روش خوشه‌بندی ایجاد گروه‌ها و طبقاتی است که تنوع درون‌گروهی آنها کمتر از تنوع و تفرق بین گروهی می‌باشد. به بیان دیگر در تجزیه خوشه‌ای معمولاً p صفت بر روی n عضو اندازه‌گیری می‌شود و بعد یک ماتریس p در n از داده‌های خام تشکیل می‌شود(فرشاد فر ۱۳۸۹: ۵۵۲).

تکنیک تحلیل خوشه‌ای بسیار وسیع و گسترده هستند. به طور خلاصه مهترین روش‌ها و تکنیک‌های خوشه‌ای مورد بررسی قرار می‌گیرند.

به طور کلی روش‌های مختلفی برای تجزیه خوشه‌ای وجود دارد:

الف- تکنیک‌ها مراتبی[۱]

ب - تکنیک‌های چند میانگینی[۲]

مراحل تجزیه‌ای خوشه‌ای مراتبی

۱- جمع آوری ماتریس داده‌ها

۲- **استاندارد کردن ماتریس داده‌ها:** در مورد استاندادر کردن داده‌ها توجه شود که اگر واحد متغیرها مورد تجزیه یکی نباشد بهتر است که اول داده‌ها به نمره z تبدیل شوند بعد تجزیه خوشه‌ای انجام شود ولی اگر داده‌ها از یک خانواده بودند شاید ضروری نباشد که داده‌ها استاندارد شوند ولی اگر استاندارد شوند باز بهتر است در متلب می‌تواند با یک تابع بسیار ساده به نام zscore داده‌های مورد بررسی را به نمره z تبدیل کرد. در واقع هدف از استاندارد کردن این است که شباهت بین متغیرهای یکسان را از بین می‌برد.

۳- **محاسبه شباهت بین متغیرهای ماتریس مورد مطالعه:** روش‌های مختلفی برای محاسبه شباهت بین متغیرها وجود دارد. این تکنیک‌ها به شرح زیر می‌باشد (عساکره ۱۳۹۱: ۲۰۱۰: ۲۰۲):

نکته: در متلب برای محاسبه این تکنیک‌ها از تابع pdist استفاده می‌شود:

ضریب متوسط فاصله اقلیدیسی[۳]: این ضریب را با d_{jk} نشان می‌دهند. از مزایای این ضریب آن است که اگر در داده‌ها مقادیری nan یا گم شده وجود داشته باشند می‌توان از آن استفاده کرد. به همین خاطر در بین روش‌ها مهمترین می‌باشد. این روش به طریق زیر محاسبه می‌شود:

$d_{jk} = \sqrt{(x_1 - x_2)^2}$	

---

[۱] - Hierarchical cluster analysis (HCA)

[۲] -K-means cluster analysis (K-MCA)

[۳] - euclidean

در این فرمول x می‌تواند هر متغیری باشد مثلا دما، بارش و رطوبت و ....

حال اگر بخواهیم تعداد بیشتری از عناصر اقلیمی را برای ایستگاه های مختلف با هم بسنجیم از رابطه ۳-۲۴ استفاده می کنیم.

$$d_{1,2} = \sqrt{\sum_{i=1}^{n}(x_{i1} - x_{i2})^2}$$

برای مثال داده‌های زیر را در نظر بگیرید:

جدول شماره--: میانگین سالانه دما، بارش و رطوبت برای سه ایستگاه خرم آباد، زنجان و مشهد

ایستگاه	دما(درجه سانتی گراد)	بارش(میلی متر)	رطوبت(درصد)
خرم آباد	۱۷/۲	۵۱۲/۳	۴۶/۷
زنجان	۱۰/۹	۲۹۸/۶۳	۵۴/۵
مشهد	۱۸/۰۵	۳۳۹/۲	۴۰

برای سه متغیر دما، بارش و رطوبت ایستگاههای مورد مطالعه نتایج زیر حاصل آمد:

$$d_{Z,KH} = \sqrt{(17.2 - 10.9)^2 + (512.3 - 298.6)^2 + (46.8 - 54.5)^2} = 213.9$$

$$d_{Z,M} = \sqrt{(18.05 - 10.9)^2 + (339.2 - 298.6)^2 + (40 - 54.5)^2} = 43.7$$

$$d_{Z,M} = \sqrt{(18.05 - 17.2)^2 + (339.2 - 512.3)^2 + (40 - 46.8)^2} = 173.2$$

نتایج به دست آمده حاکی از کم ترین فاصله اقلیدسی سه عنصر مورد بررسی در ایستگاه زنجان و مشهد می باشد.

برای محاسبه این روش که مهمترین روش هم می‌باشد به طریق زیر عمل خواهیم کرد:

```
>> nam=pdist(dda,'euclidean')
nam =
  213.9051  173.2317   43.6726
```

ضریب تفاوت شکل: این ضریب را با z_{jk} نشان می‌دهند. برای محاسبه این ضریب از فرمول زیر استفاده می‌کنیم:

$$z_{jk} = [(\frac{n}{n-1}) \times (\Sigma d_{jk1}^2 - \Sigma d_{jk2}^2)]^2$$

در فرمول فوق n طول دوره آماری و d_{jk1}^2 به ظریق زیر بدست می‌آید:

$$d_{jk1}^2 = (\frac{1}{n^2})(\sum_{i=1}^{n} d_{i\,j} - \sum_{i=1}^{n} d_{i\,k})^2$$

مثال: با توجه به جدول – فاصله ایستگاه‌ها به شرح زیر است:

خرم آباد - زنجان بر اساس ضریب تفاوت شکل

$$q^2{}_{Z,KH} = \left(\frac{1}{3^2}\right)([17.2 + 512.3 + 46.8] - [10.9 + 298.6 + 54.5])^2$$

$$q^2{}_{Z,KH} = 5007.9$$
$$Z_{Z.KH} = \sqrt{\left[\frac{3}{3-1}\right](45753.21 - 5007.9)} = 247.2$$

محاسبه فاصله زنجان - مشهد بر اساس ضریب تفاوت شکل

$$q^2{}_{Z,M} = \left(\frac{1}{3^2}\right)([10.9 + 298.6 + 54.5] - [18.05 + 339.2 + 40])^2$$
$$Z_{Z.M} = \sqrt{\left[\frac{3}{3-1}\right](1909.69 - 122.8)} = 51.8$$

محاسبه فاصله مشهد - خرم آباد بر اساس ضریب تفاوت شکل:

$$q^2{}_{M,KH} = \left(\frac{1}{3^2}\right)([17.2 + 512.3 + 46.8] - [18.05 + 339.2 + 40])^2$$
$$Z_{M.KH} = \sqrt{[\frac{3}{3-1}](29998.24 - 3562.1)} = 199.1$$

با توجه به محاسبات انجام شده کمترین ضریب تفاوت مربوط به زنجان – مشهد می باشد به این معنا که این دو ایستگاه از نظر اقلیمی نسبت به سایر ایستگاههای سنجیده شده تفاوت کم تری دارند.

ضریب کسینوس[۴]: این ضریب را با c_{jk} نمایش می‌دهند و دامنه آن بین ± 1 می‌باشد. اگر این مقدار یک باشد یعنی شباهت دارای حداکثری می‌باشد. اگر منفی یک باشد یعنی مقدار شباهت حداقل می‌باشد. این نمایه به طریق زیر محاسبه می‌شود:

$$C_{jk} = \frac{\sum_{i=1}^{n} x_{ij} x_{ik}}{\sqrt{\left(\sum_{i=1}^{n} x_{ij}{}^2\right)\left(\sum_{i=1}^{n} x_{ik}{}^2\right)}}$$

مثال برای خرم آباد و زنجان

$$\sum_{i=1}^{3} x_{KH} x_Z = (17.2 \times 10.9) + (46.8 \times 54.5) + (512.3 \times 298.6) = 155710.8$$
$$\sum_{i=1}^{3} x_{KH}{}^2 = (17.2)^2 + (46.8)^2 + (512.3)^2 = 264937.37$$
$$\sum_{i=1}^{3} x_Z{}^2 = (10.9)^2 + (54.5)^2 + (298.6)^2 = 92251.02$$
$$C_{\mathrm{KH,Z}} = \frac{155710.86}{\sqrt{264937.37 \times 92251.02}} = 0.004$$

میزان به دست آمده شباهت حداکثر بین دو ایستگاه خرم آباد و زنجان را نشان می دهد این در حالی است که مقادیر مربوط به عناصر دو ایستگاه مشابه نمی باشند. محاسبه این ضریب برای دو ایستگاه خرم آباد و مشهد

[۴] - cosine

نتیجه ۰/۹۹۹ را به دست داد که شباهت حداکثر را نشان می دهد.

ساختار این روش در متلب:

```
>> nam=pdist(dda,'cosine')
nam =
   0.0040    0.0005    0.0021
```

ضریب متریک کانبرا[۵]: این ضریب را با a_{jk} نشان می‌دهند در دامنه بین صفر و یک قرار دارد. هر چه این مقدار به صفر نزدیک تر باشد نشان از شباهت زیاد ایستگاه مورد بررسی می‌باشد. میزان این ضریب از -- قابل محاسبه است.

$$a_{jk} = \left(\frac{1}{n}\right)\sum_{i=1}^{n}\frac{|x_{ij}-x_{jk}|}{(x_{ij}+x_{jk})} - ۳۰)$$

به محاسبات دقت کنید:

$$a_{Z,KH} = \left(\frac{1}{3}\right)\sum_{i=1}^{n}\frac{|17.2-10.9|}{(17.2+10.9)}+\frac{|512.3-298.6|}{(512.3+298.6)}+\frac{|46.8-54.5|}{(46.8+54.5)} = 0.188$$

میزان به دست آمده شباهت زیاد دو ایستگاه را نشان می دهد این میزان نسبت به صفر سنجیده می شود و هر چه رقم حاصله به صفر نزدیک تر باشد شباهت بیشتر است.

ضریب بری – کرتیس: محاسبه این ضریب از رابطه -- امکان پذیر می باشد. میزان این ضریب بین صفر و یک به دست می آید و هر چه به صفر نزدیک تر باشد شباهت بیشتر خواهد بود.

$$b_{jk} = \frac{\sum_{i=1}^{n}|x_{ij}-x_{jk}|}{\sum_{i=1}^{n}(x_{ij}+x_{jk})}$$

مثال

$$b_{Z,K} = \frac{|17.2-10.9|+|512.3-298.6|+|46.8-54.5|}{(17.2+10.9)+(512.3+298.6)+(46.8+54.5)} = 0.242$$

همانگونه که ذکر شده است میزان صفر برای این ضریب حداکثر شباهت است و با فاصله از آن میزان شباهت کاسته می شود بنابر این شباهت بین دو ایستگاه زنجان و خرم آباد نسبتاً زیاد است.

فاصله مینسکاوسکی[۶](توانی): برای سرشکن کردن تأثیر بزرگی برخی متغیرها که تفاوت فاحشی با بقیه دارند از رابطه -- به محاسبه این ضریب می پردازیم. در اینجا m فاصله بلوک شهری (منهاتن) است که آن را برابر ۱در نظر می گیرند.

---

[۵] - camberra metric coefficent

[۶] - minkowski

$$d_{ij} = \left[\sum_{k=1}^{p} \left|x_{ik} - x_{jk}\right|^{m}\right]^{\frac{1}{m}} \quad d_{Z.KH}$$
$$= |17.2 - 10.9| + |46.8 - 54.5| + |512.3 - 298.6| = 173.67$$

میزان فاصله محاسبه شده برای خصیصه های اقلیمی خرم آباد و زنجان در فاصله اقلیدسی ۲۱۳/۹ بوده در حالی که این میزان برای فاصله توانی ۲۲۷/۷ می باشد، این میزان پس از سر شکن کردن اعداد بزرگ حاصل آمده است.

در متلب:

```
>> nam=pdist(dda,'minkowski')
nam =
  213.9051 173.2317   43.6726
```

۴- استفاده از روش کلاستر برای دسته بندی متغیرها:

هدف از تکنیک های دسته بندی آن است که مجموعه‌ای از داده‌ها در گروه‌ها یا دسته های مجزا قرار گیرند. این تکنیک ها را می توان در تکنیک های مراتبی[۷]، تکنیک های اپتمی کردن[۸]، تکنیک های دسته‌ای[۹]، تکنیک های تراکمی[۱۰] و گرافیکی[۱۱] طبقه بندی کرد. تنیک های مراتبی خود به دو روش تجمعی[۱۲] و تقسیمی[۱۳] قابل بررسی‌اند. روش‌های تجمعی یا تراکمی خود به انواعی دیگر که در واقع به تکنیک های مهم دسته بندی هم معروف هستند تقسیم می‌شوند:

نکته: برای محاسبه این تکنیک ها در متلب از تابع linkage استفاده می‌کنیم.

۱-روش همبستگی منفرد یا نزدیک ترین همسایه ('single '): در این روش از حداقل فاصله بین متغیرها استفاده می‌شود. به عبارتی دیگر در این روش فاصله بین کلاسترها را فاصله بین نزدیکترین اعضا تشکیل می‌دهد. نزدیک فاصله به صورت زیر تعریف می‌شود:

$$d_{h\,k} = \min(d_{ih}, d_{jk})$$

این فرمول این را بیان می‌کند که در بین متغرها مورد بررسی حداقل را انتخاب و بر این مبنا کلاس بندی انجام می‌شود. البته این روش را می‌توان با فرمول زیر هم حساب کرد:

$$d_{h\,k} = \frac{1}{2}(h_{hi}) + \frac{1}{2}(h_{hj}) - \frac{1}{2}\left|h_{hi} - h_{hj}\right|$$

بنابراین چه از این فرمول چه از فرمول قبلی استفاده کنیم نتایج یکی می‌باشد. برای مثال دسته بندی داده‌ها به روش نزدیک ترین همسایه بر روی فراوانی یخبندان در استان یاسوج در ماه‌های آذر به شرح زیر می‌باشد:

---

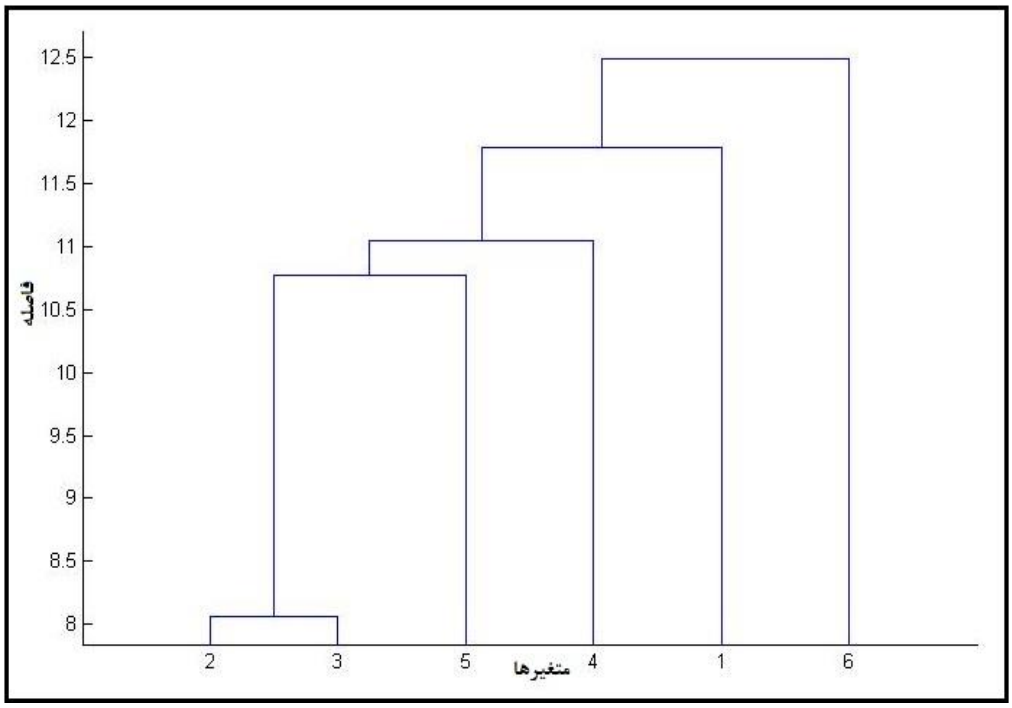
[7] - Hierarchical
[8] -Optimization Techniques
[9] - Clumping Techniques
[10] -Density Techniques
[11] - Graphical Techniques
[12] -Agglomerative Method
[13] - Divisive Method

```
x =
   26   14   16   13
   12    8   14   21
   12   14   19   23
   22   10   11   24
   21   15   24   20
   23   25   11   12
```

با انجام محاسبات مربوط به این روش ماتریس فاصله ها به شرح زیر می‌باشد. البه لازم به ذکر است که برای شباهت از فاصله اقلیدیسی استفاده شده است:

```
>> Z =linkage(nn,'single','euclidean');
Z =
        1        2      3.873
        4        5     4.6904
        3        7     8.8318
        6        8      11.18
```

برای مثال ۱ و ۲ دارای فاصله‌ای ۳/۸۷۳ می‌باشد که بیشترین شباهت را برای قرار گیری در دیک دسته دارند نتایج دندروگرام حاصل از نزدیک ترین همسایه به شرح زیر است:



شکل—نتایج خوشه‌ای به روش نزدیک ترین همسایه

۲- **روش همبستگی کامل یا دورترین همسایه (complete linkage')**: این روش درست عکس روش قبل است یعنی دور ترین فاصله را در نظر می‌گیرد. فاصله بین دست‌ها در اینجا با فرمول زیر قابل محاسبه می‌باشد:

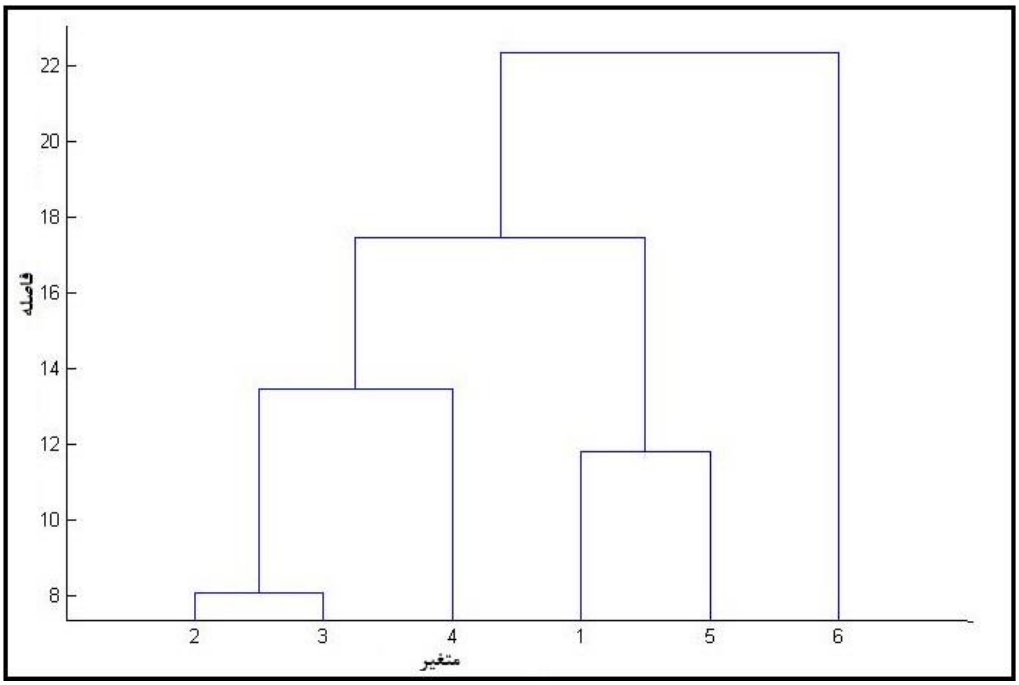
$d_{h\,k} = \max(d_{ih}, d_{jk})$	

این فرمول این را بیان می‌کند که در بین متغرها مورد بررسی حداکثر را انتخاب و بر این مبنا کلاس بندی انجام می‌شود. این محاسبات را می‌توان با فرمول زیر هم حساب کرد:

$d_{h\,k} = \frac{1}{2}(h_{hi}) + \frac{1}{2}(h_{hj}) + \frac{1}{2}\lvert h_{hi} - h_{hj}\rvert$	

برای داده های قبلی این بار برای دسته بندی از روش دورترین همسایه استفاده شده است که نتایج زیر حاصل گشته است:

```
>> Z =linkage(nn,'complete','euclidean');
Z =
   2.0000    3.0000    8.0623
   1.0000    5.0000   11.7898
   4.0000    7.0000   13.4536
   8.0000    9.0000   17.4642
   6.0000   10.0000   22.3607
```



شکل—نتایج خوشه‌ای به روش دورترین همسایه

۳- **روش خوشه‌ای همبستگی متوسط (Unweight 'average' distance or UPGMA)**: در این روش دو گروه وقتی با یک دیگر ادغام می‌شود که فاصله متوسط بین آنها به اندازه کافی کوچک باشد. برای محاسبه این روش به طریق زیر عمل خواهیم کرد:

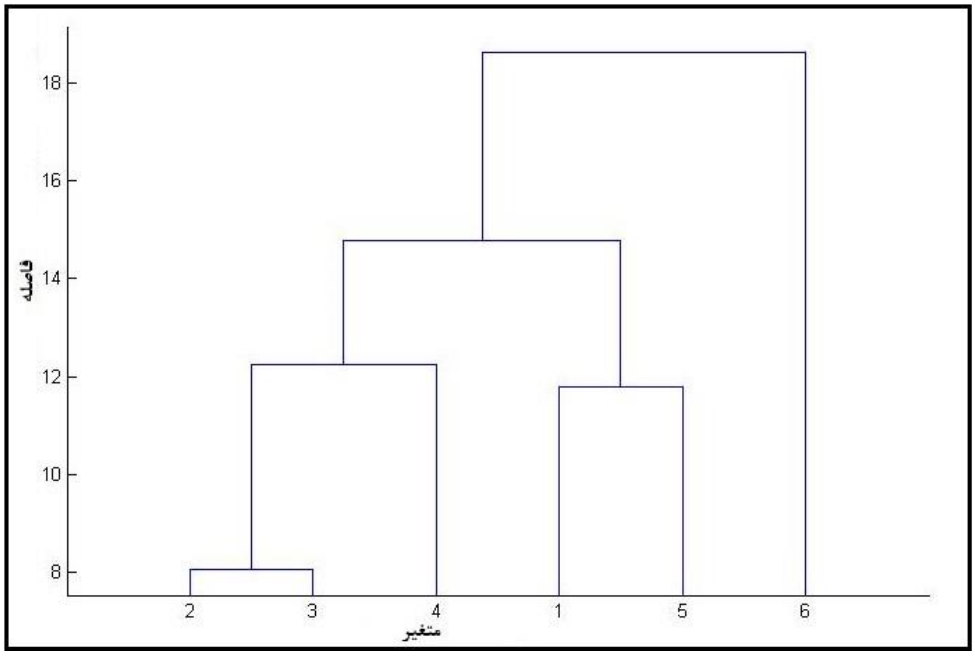
$d_{h\,k} = \frac{n_i}{n_k} \times (h_{hj}) + \frac{n_j}{n_k} \times (h_{hi})$	

در این فرمول از طریق این رابطه $n_k = (n_i - n_j)$ به دست می‌آید.

ماتریس فاصله

```
>> Z =linkage (nn, ‘average’,'euclidean')
Z =
   2.0000    3.0000    8.0623
   1.0000    5.0000   11.7898
   4.0000    7.0000   12.2495
   8.0000    9.0000   14.7702
   6.0000   10.0000   18.6215
```

اگر دقت کنید در هر مرحله فاصله دسته بندی متغیرها تفاوت‌های قابل محسوسی را نشان می‌دهند.



شکل—نتایج خوشه‌ای به روش همبستگی متوسط

۴- روش خوشه‌ای محوری ('centroid'):

در این روش فاصله بین متغیرها بر مبنای میانگین است و از فاصله مربع استفاده می‌شود سپس آنهای که کوچکترین فاصله را دارند ابتدا ادغام می‌شوند. البته محققین از این روش به علت نقصی که در ارایه دندروگرام می‌دهد زیاد استفاده نمی‌شود. دلیل نقص آن این است که مقادیری که در این دسته ادغام می‌شود به جای اینکه افزایش یابد کاهش می‌یابد و تفسیر آن را با مشکل مواجه می‌کند (فرشادفر ۱۳۸۹: ۵۸۷). این روش به طریق زیر محاسبه می‌شود:

$$d_{h\,k} = \frac{n_i}{n_k} \times \left(h_{hj}\right) + \frac{n_j}{n_k} \times \left(h_{hi}\right) - \frac{n_i\, n_j}{n_k^2} \times \left(h_{ij}\right)$$

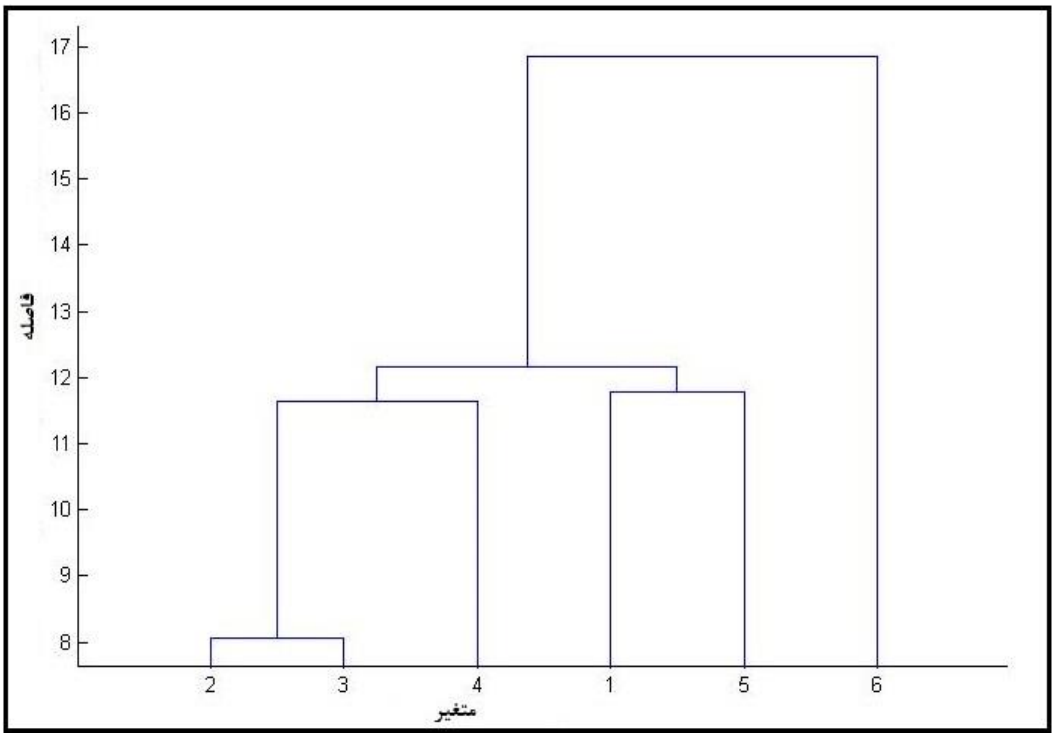
در این فرمول از طریق این رابطه $n_k = (n_i - n_j)$ به دست می‌آید.

ماتریس فواصل متغیرها بر اساس این روش به شرح زیر است:

```
>> Z =linkage(nn,'centroid' ,'euclidean')
Z =
   2.0000    3.0000    8.0623
```

```
   4.0000    7.0000   11.6297
   1.0000    5.0000   11.7898
   8.0000    9.0000   12.1598
   6.0000   10.0000   16.8547
```

همان طوری که مشاهده می‌شود نتایج این دسته بندی با دسته بندی به روش میانگین تفاوت بسیار اندکی دارند که این به دلیل اینکه در هر دو روش مرکز داده‌ها را مد نظر دارند می‌باشد.



شکل—نتایج خوشه‌ای به روش

۵- **روش خوشه‌ای همبستگی میانه ('median')**: زمانی که از روش دستبه بندی محوری برای ادغام دو گروه استفاده می‌شود که از نظر اندازه متفاوت هستند محور کلاستر جدید به دسته یزرگتر میل می‌کند که این خود باعث در نظر نگرفتن دسته کوچک می‌شود. برای حل این مشکل روش همبستگی میانه به شرح زیر استفاده خواهیم کرد:

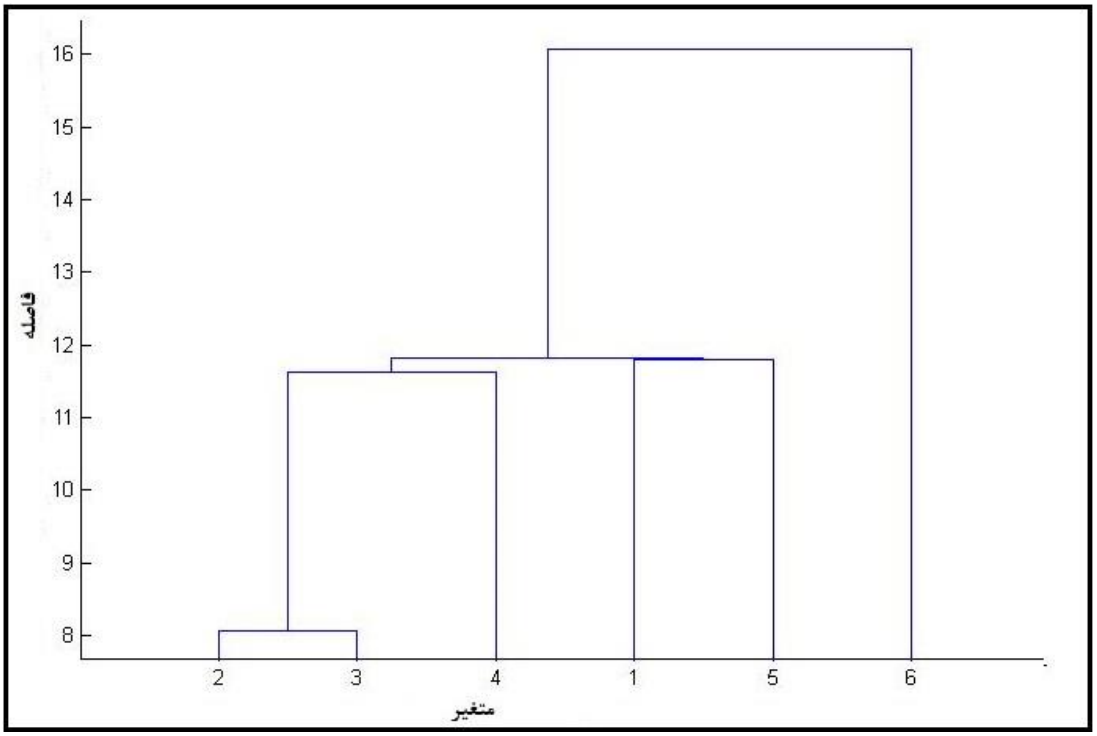
$d_{h\,k} = \frac{1}{2}(h_{hi}) + \frac{1}{2}(h_{hj}) - \frac{1}{4}(h_{ij})$	

ماتریس فاصله‌ها

```
>> Z =linkage(nn,'median' ,'euclidean');
Z =
   2.0000    3.0000    8.0623
   4.0000    7.0000   11.6297
   1.0000    5.0000   11.7898
   8.0000    9.0000   11.8137
```

6.0000 10.0000 16.0745

همان طوری که اشاره شد در این مثال‌ها روش دسته بندی در هر مرحله فرق می‌کند اما برای شباهت و تفاوت از روش اقلیدیسی برای مثال استفاده می‌کنیم که نسبت به سایر روش‌ها ارجحیت بیشتری هم دارد.



شکل—نتایج خوشه‌ای به روش همبستگی میانه

۶- روش خوشه‌ای حداقل وارد ('ward'):

در بین روش‌های یاده شده روش وارد از اهمیت به سزایی در اقلیم برخوردار است. این روش از گروه روش های تجمعی است و در سال ۱۹۳۶ توسط وارد پیشنهاد شده است. این روش به این صورت است که در هر مرحله از تجزیه، کمبود اطلاعاتی را که در اثر دسته بندی افراد در دسته ها به وجود می آید را می توان توسط کل مجموع مربعات انحرافات هر نقطه از میانگین دسته ای که به آن تعلق دارد به دست آورد. در هر مرحله دو دسته ای که ادغام آنها سبب افزایش مجموع مربعات اشتباه[۱۴] به میزان حداقل شود با هم ترکیب می شوند. افرادی که در یک جفت از دسته ها دارای حداقل مجموع مربعات اشتباه(ESS) هستند در یک دسته قرار می گیرند. برای این کار باید مقدارESS را به دست آورد که از طریق فرمول زیر به دست می آید.

$$E.S.S = \sum_{j=1}^{k}\left[\sum_{i=1}^{n_j} x_{ij}^2 - \frac{1}{n_j}\left[\sum_{i=1}^{n_j} x_{ij}\right]^2\right]$$

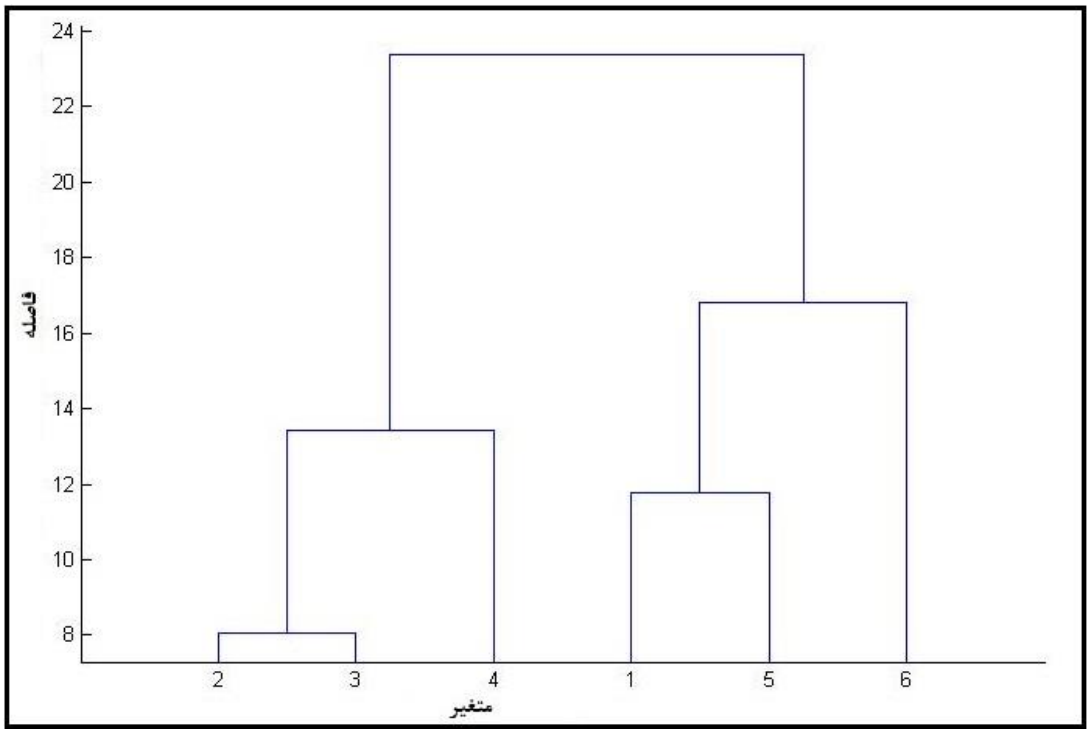
در این فرمول x_{ij} نمره فرد iام در دسته jام، و k تعداد کل دسته‌ها در هر مرحله، از طرف دیگر n_j تعداد افراد در هر دسته jام است. این مجموع مربعات اشتباه را شاخص مجموع مربعات یا واریانس گویند (فرشادفر ۱۳۸۹،

1. Error sum of squares

۵۹۲).

ماتریس فاصله‌ها به روش وارد:

```
>> Z =linkage(nn,'ward' ,'euclidean')
Z =
   2.0000    3.0000    8.0623
   1.0000    5.0000   11.7898
   4.0000    7.0000   13.4288
   6.0000    8.0000   16.8028
   9.0000   10.0000   23.3666
```



شکل—نتایج خوشه‌ای به روش وارد

دستورات تحلیل خوشه‌ای در متلب

حالا اگر بخواهیم به طور خلاصه مهترین دستورات تحلیل خوشه‌ای را در متلب بنویسیم دو روش دستورات وجود دارد که در هر دو روش یکی حاص می گردد:

روش اول: در این روش ابتدا محاسبه شباهت بین متغیرهای ماتریس مورد مطالعه با دستور pdist می‌باشد که به طریق زیر عمل خواهیم کرد. البته این نکته را هم به یاد داشته باشید که پیش فرض این دستور فاصله اقلیدیسی می‌باشد:

```
>> Y=pdist(x);
```

اگر این دستور را به تنهایی بکار ببرید متلب به طور پیش فرض شباهت و تفاوت بین متغیرها را بر اساس فاصله

اقلیدیسی حساب می کند ولی اگر خواستید روش دیگری را انتخاب کنید به طریق زیر عمل خواهیم کرد:

مرحله اول:

```
>> Y=pdist(x, metod);
```

در اینجا به جای metod می‌توانید روش‌های که در بالا توضیح داده شده قرار دهید. برای مثال می‌توانید از روش 'spearman' یا روش همبستگی، 'cosine' یا روش کسینوس، 'chebychev' یا چبی چوف و 'minkowski' یا مینسکاوسکی و ... که در مطالب قبل شرح داده شد استفاده کرد.

مرحله دوم: در این مرحله باید روش دسته بندی را با استفاده از تابع linkage حساب کرد. که روش‌ها و تکنیک‌ها آن در مطالب فوق به طور مفصل شرح داده شده است. برای مثال به طریق زیر عمل خواهیم کرد:

```
>> Z = linkage(Y, ‘ward’);
```

مرحله سوم: در این مرحله با استفاده از تابع dendrogram شکل خوشه‌ها را ترسی می‌کنیم. که به طریق زیر عمل می‌کنیم:

```
>> dendrogram(Z,Rowe);
```

مرحله چهارم: در این مرحله تعداد کلاس‌ها مناسب را با تابع cluster استخراج خواهیم کرد:

```
>> clusterrcluster(Z, Rowe);
```

روش دوم: در این روش در متلب می‌توان با یک دستور همزمان هم شباهت و تفاوت را و هم روش‌های دسته بندی متغیرها را محاسب کرد شکل کلی ساختار این دستور به طریق زیر است.

مرحله اول:

```
>> Z = linkage(Y,metod, pdist);
```

حالا می توانیم هم روش شباهت و تفاوت و هم روش ادغام را مطابق ساختار دستور بالا نوشت به دستور زیر توجه کنید.

```
Z =linkage(nn,'average' ,'euclidean')
```

مرحله دوم: در این مرحله با استفاده از تابع dendrogram شکل خوشه‌ها را ترسی می‌کنیم. که به طریق زیر عمل می‌کنیم:

```
>> dendrogram(Z,Rowe);
```

مرحله سوم: در این مرحله تعداد کلاس‌ها مناسب را با تابع cluster استخراج خواهیم کرد:

```
>> cluster=cluster(Z, Rowe);
```

مثال: داده‌های پنج سال شیراز را در نظر بگیرید:

ژوئن	می	آوریل	مارس	فوریه	ژانویه	سال

1951	7/3	7/9	10/9	16/4	21/6	26/6
1952	6/4	9/3	12/7	16/8	22	26/3
1953	8/4	9/4	11/4	16/7	21/9	26/7
1954	8/9	8/1	12/1	16/2	22/7	26/8
1955	6/4	9/2	12/6	16/9	21/4	26/1
1956	3/6	9	11/8	15/4	21/8	27
1957	4/5	6/3	10/9	13/2	18/6	25/1
1958	6/4	7/6	13/1	18	22	25/3
1959	4/9	5/1	11/4	18/2	21/5	26/4

همان طوری که اشاره شد بهترین روش محاسبه شباهت و تفاوت در تحلیل خوشه‌ای فاصله اقلیدیسی و بهرین روش دسته بندی یا ادغام روش وارد می‌باشد به همین خاطر در این مثال از این دو روش استفاده می‌کنیم. ترتیب وارد کردن دستورات که از روش دوم برای این مثال استفاده شده است:

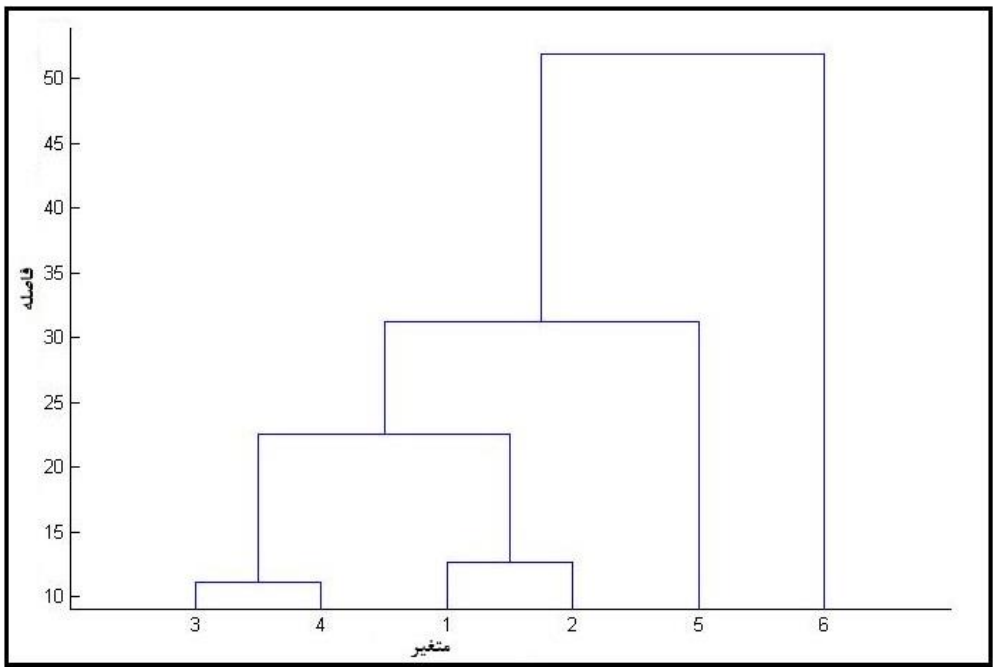
مرحله اول: ماتریس فاصله ها بر اساس روش اقلیدیسی و روش دسته بندی وارد

```
>> Z =linkage(data,'ward' ,'euclidean')
Z =
     1     3     3.10
     5     6     5.78
     7    11     7.30
     9    10     8.78
     4    13    14.27
     2     8    19.66
    14    15    32.44
    12    16    43.50
```

در ماتریس بالا فاصله‌های هر گروه در ستون سوم آورده شده است. برای مثال ۱ و ۳ در فاصله ۳/۱۰ که کمترین فاصله می‌باشد بیتر شباهت را دارد. به عبارتی از اولویت بیشتری برای تشکیل یک کلاستر بر خوردار می‌باشد.

مرحله دوم: با استفاده از دستور زیر نمودار درختی این فاصله‌ها را نمایش می‌دهیم:

```
>> dendrogram(Z);
```



مرحله سوم: تعدا خوشه‌ها را با دستور زیر بیان می‌کنیم:

```
>> vv=cluster(Z,2)
vv =
    2
    2
    2
    2
    2
    1
```

در این دستور به جای ۲ می توانید تعداد کلاسترهای دلخواه را وارد کنید. برای مثال می توانید ۳ یا ۴ قرار دهید آنگاه آنها را در چهار گروه یا سه گروه تقسیم می‌کند. حلا برای تعیین تعداد گروه‌های مناسب می‌توانید از آزمون t که شرح داده شده‌اند بهترین گروه را انتخاب کرد. در واقع گروهی مناسب می‌باشد که کمترین درصد خطا را داشته باشد. ساختار دستور برای کلاس بندی‌های دیگر:

```
 >> vv(:,1)=cluster(Z,2);
>> vv(:,2)=cluster(Z,3);
>> vv(:,3)=cluster(Z,4);
>> vv(:,4)=cluster(Z,5)
vv =
    2    2    2    1
    2    2    2    2
    2    2    1    3
    2    2    1    3
    2    1    3    4
```

1 3 4 5

مثال ۲: چگونگی مطالعه بر روی خوشه‌ها

معمولا این سوال برای دانشجویان پیش می‌آید که چگونه می‌توان با حجم داده‌های وسیع ویژگی خوشه های یک عنصر اقلیمی را مورد بررسی قرار داد. در این مثال روی چند ایستگاه فرضی با متغیرهای فرضی این کار را در ادامه انجام خواهیم داد. داده های زیر را در نظر بگیرید اسم این ماتریس در محیط متلب فرض کنید data باشد:

جدول—عناصر اقلیمی برای ایسگاه‌های فرضی

ایستگاه	دما	بارش	رطوبت	باد	فشار
A	16.60	17.70	16.27	18.66	20.93
B	16.27	18.20	17.21	19.35	22.06
C	16.35	17.74	16.71	19.02	21.84
D	17.44	18.75	16.86	18.59	21.01
E	17.19	18.74	17.59	19.72	22.48
F	14.58	16.44	15.87	18.94	22.38
G	15.76	17.81	16.93	19.46	22.09
H	16.11	17.86	16.80	19.24	21.85
K	14.80	15.95	14.27	16.60	20.10
N	13.58	15.65	15.42	18.72	22.09
M	15.89	18.03	17.50	20.11	22.85
X	15.42	16.91	16.24	19.02	22.04
Z	14.35	16.43	15.96	18.25	20.58
Q	15.28	17.18	16.30	19.00	21.69
R	15.70	17.23	16.09	18.48	21.37
T	15.64	17.10	15.87	18.61	21.70
Y	14.88	15.57	13.88	16.63	20.30
U	15.38	16.29	14.85	17.50	20.45
P	14.56	15.06	13.44	16.08	19.25
L	13.52	14.61	13.49	16.46	19.80

حل

در ابتدا باید داده هارا با استفاده از دستور زیر استاندارد کنیم:

```
>> Zsc=zscore (data);
```

محاسبه فاصله اقلیدیسی و روش ادغام:

```
>> Z =linkage (Zsc,'ward' ,'euclidean');
```

ترسیم دندروگرام

```
>> dendrogram(Z,20);
```

یا یکی از دو دستور زیر را برای ترسیم دندروگرام انتخاب کنید:

```
>> H = dendrogram (Z,'Orientation','left','ColorThreshold','default');
```

```
>> H = dendrogram (Z,'Orientation','right','ColorThreshold','default');
```

با اجرای دستور دندروگرام شکل --- حاصل خواهد شد.

در مرحله بعد تعدا کلاسترها را باید مشخص کنیم مثلا دو گروهه، سه گروهه و .... که برای این کار از تابع زیر استفاده خواهید کرد:

```
>> Clust(:,1)=cluster(Z,2);
>> Clust(:,2)=cluster(Z,3);
>> Clust(:,3)=cluster(Z,4);
>> Clust(:,4)=cluster(Z,5);
```

اگر دستورات بالا را اجرا کنیم ۴ ستون ایجاد می کند به طوری که ستون اول تعداد ایستگاهها را به دو خوشه، ستون دوم به سه خوشه و .... تقسیم می کند. نتایج حاصل از دستورات بالا را متغیرهای اصلی ک بار دیگر نشان خواهیم داد:

متغیرهای اصلی					گروه‌ها				
فشار	**باد**	**رطوبت**	**بارش**	**دما**	**پنجم**	**چهارم**	**سوم**	**دوم**	**ایستگاه**
20.93	18.66	16.27	17.70	16.60	3	1	2	2	A
22.06	19.35	17.21	18.20	16.27	4	2	2	2	B
21.84	19.02	16.71	17.74	16.35	4	2	2	2	C
21.01	18.59	16.86	18.75	17.44	3	1	2	2	D
22.48	19.72	17.59	18.74	17.19	4	2	2	2	E
22.38	18.94	15.87	16.44	14.58	2	3	1	2	F
22.09	19.46	16.93	17.81	15.76	4	2	2	2	G
21.85	19.24	16.80	17.86	16.11	4	2	2	2	H
20.10	16.60	14.27	15.95	14.80	5	4	3	1	K
22.09	18.72	15.42	15.65	13.58	2	3	1	2	N
22.85	20.11	17.50	18.03	15.89	4	2	2	2	M
22.04	19.02	16.24	16.91	15.42	1	3	1	2	X
20.58	18.25	15.96	16.43	14.35	2	3	1	2	Z
21.69	19.00	16.30	17.18	15.28	1	3	1	2	Q
21.37	18.48	16.09	17.23	15.70	1	3	1	2	R
21.70	18.61	15.87	17.10	15.64	1	3	1	2	T
20.30	16.63	13.88	15.57	14.88	5	4	3	1	Y
20.45	17.50	14.85	16.29	15.38	5	4	3	1	U
19.25	16.08	13.44	15.06	14.56	5	4	3	1	P
19.80	16.46	13.49	14.61	13.52	5	4	3	1	L

در مرحله بعد سوال اینجا پیش می‌آید که چند گروه مناسب است. یا ایستگاه‌ها ما بر اساس این متغیرها در چند گروه قابل تقسیم هستند؟ محل برش را در شکل دندروگرام از کجا بزنیم؟.

کارگاه برنامه نویسی در متلب برای دانشجویان خوارزمی جلسه پنجم و ششم- مهدی دوستکامیان

همان طوری که هم قبلا اشاره شد چند عامل خیلی موثر می‌باشد:

۱- تجربه و شناخت از منطقه خیلی کمک می کند.

۲- استخراج آمار توصیفی تمام گروه‌ها: در آمار توصیفی بیشتر به چند فراسنج حساس باشید. اولا به میانگین عنصرها در گروه‌ها و مقایسه آنها با همدیگر. چنانچه در یک از گروه‌ه میانگین ها بسیار به هم نزدیک باشد نمی توان آن گروه را به عنوان محل برش انتخاب کرد. دوم واریانس و انحراف معیار که هر چه که در گروه مورد نظر اختلاف داشته باشند برای برش دندروگرام شایسته تر می باشد. این روش‌ها تقریب روش توصیفی هستند و می توانند سرنخ دهند.

۳- اما برای برش دندرو گرام باید از روش های و آزمون های آماری به خصوص آزمون های اختلاف میانگین نظیر t که در فصل آزمون‌ها به طور مفصل شرح داده شده اند استفاده کرد. خالا اینجا سوال پیش می‌آی که چطوری استفاده کنیم. برای این کار مثلا دو گروه میانگین گروه اول را با میانگین گروه دوم همان گروه مقایسه می کنیم. در اینجا چون پنج تا متغیر داریم باید باید برای هر مقایه پنج آماره دو به دو که میشه ۱۰ آماره را حساب کنیم مثلا پنج متغیره گروه اول با پنج متغیره گروه دوم به همین ترتیب برای سایر گروه‌ها این کار را باید انجام دهیم سپس تعداد خطاها را بر تعداد کل آماره ها تقسیم می کنیم آن گروهی که کمترین خطا را داشته باشد به عنوان گروه بهینه جهت برش دندروگرام بکار می رود. همان طوری که مشاهده می شود در اینجا سه گروهه کمترین درصد خطا را داشته است به همین خاطر انتخاب شده است:

گروه	گروه	تعداد خطا ها	درصد خطا
گروه دو	گروه یک	۲	$RS = 2/10 = 0.2$

*مجموع مقدار خطا برای گروه دوم ۲/ حاصل شده است.

سه گروه:

گروه	گروه	تعداد خطا ها	درصد خطا
گروه دو	گروه یک	۲	$RS = 2/10 = 0.2$
گروه سه	گروه یک	۰	$RS = 0/10 = 0$
گروه سه	گروه دو	۰	$RS = 0/10 = 0$

*مجموع مقدار خطا برای گروه سوم ۲/ حاصل شده است.

چهار گروه:

گروه	گروه	تعداد خطا ها	درصد خطا
گروه چهار	گروه یک	۱	$RS = 1/10 = 0.1$
گروه سه	گروه یک	۷	$RS = 7/10 = 0.7$
گروه دو	گروه یک	۶	$RS = 6/10 = 0.6$
گروه چهار	گروه دو	۵	$RS = 5/10 = 0.5$
گروه سه	گروه دو	۴	$RS = 4/10 = 0.4$

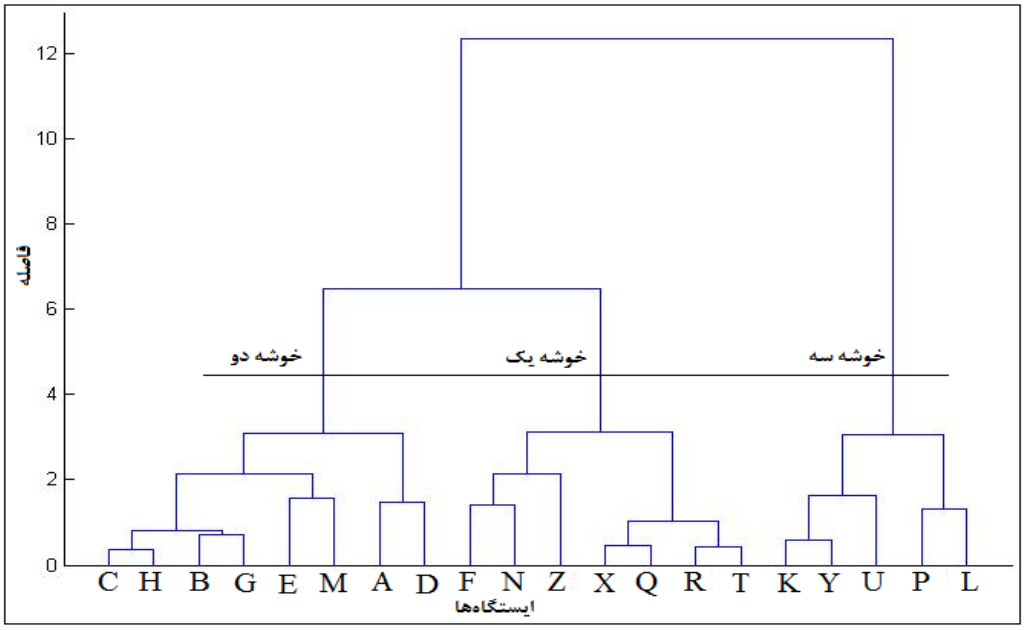
گروه چهار	گروه سه	۳	$RS = 3/10 = 0.3$

*مجموع مقدار خطا برای گروه چهارم ۲/۶ حاصل شده است.

پنج گروه:

گروه	گروه	تعداد خطا ها	درصد خطا
گروه پنج	گروه یک	۳	$RS = 3/10 = 0.3$
گروه چهار	گروه یک	۴	$RS = 4/10 = 0.4$
گروه سه	گروه یک	۶	$RS = 6/10 = 0.6$
گروه دو	گروه یک	۵	$RS = 5/10 = 0.5$
گروه پنجم	گروه دو	۷	$RS = 7/10 = 0.7$
گروه چهار	گروه دو	۳	$RS = 3/10 = 0.3$
گروه سه	گروه دو	۵	$RS = 5/10 = 0.5$
گروه پنجم	گروه سه	۶	$RS = 6/10 = 0.6$
گروه چهار	گروه سه	۵	$RS = 5/10 = 0.5$
گروه پنجم	گروه چهار	۷	$RS = 7/10 = 0.7$

*مجموع مقدار خطا برای گروه پنجم ۵/۱ حاصل شده است.



شکل -- نتایج حاصل از تحلیل خوشه‌ای به روش وارد

همان طوری که اشاره شده است سه گروه مناسب بوده است. که در زیر مشخصات توصیفی هر پنج گروه آورده شده است. اگر دقت کنید بر اساس همین امار توصیفی هم می توانستیم سه گروهه را به راحتی انتخاب کنیم. اگر ایستگا مختصات داشته باشد می توانید خوشه‌ها خودر را در قالب یک نقشه مطابق آنچه که در مثال آب قابل بارش در ادامه آورده شده است نمایش دهید.

کارگاه برنامه نویسی در متلب برای دانشجویان خوارزمی جلسه پنجم و ششم- مهدی دوستکامیان