تحلیل خوشه ای

مقدمه

در این قسمت ابتدا چند تعریف بیان می کنیم و در ادامه به جزئیات این تعاریف و کاربردهای تحلیل خوشه ای در علوم مختلف می پردازیم و نیز با مشکلاتی که در تحلیل خوشه ای مواجه هستیم، اشاره ای داریم و انتظار ما از انجام آن چیست. دو ابزار مهم خوشه بندی، ماتریس الگو و ماتریس مشابهت را معرفی می کنیم.

همچنین در این قسمت، درخت انواع رده بندی را معرفی می کنیم، و چون خوشه بندی خود زیر مجموعه ای از رده بندی است، شاخه ای که (خوشه بندی بدون ناظر) در این فصل به تفسیر و توضیح آن می پردازیم را مشخص می کنیم و در نتیجه الگوریتم جانسون برای خوشه بندی سلسه مراتبی تجمعی (و انواع این خوشه بندی) و همچنین الگوریتم خوشه بندی افرازی (و یک نوع مهم آن به نام K-میانگین) را بیان می کنیم.

خوشه: اوریت (۱۹۷۴) چندین تعریف برای خوشه بیان کرده است.

۱. هر خوشه مجموعه ای از نمونه های متشابه است و نمونه های خوشه های مختلف نامتشابه اند.
۲. خوشه اجتماعی از نقاط در فضای نمونه است، بطوری که فاصله بین هر دو نقطه در یک خوشه، از فاصله بین دو نقطه ای که در یک خوشه نیستند کمتر می باشد.
۳. خوشه را می توان ناحیه ای همبند از فضای چند بعدی شامل نقاطی با چگالی نسبتاً زیاد توصیف کرد.

خوشه بندی فرآیند دسته بندی مجموعه ای از n نمونه به خوشه ها است، نمونه های متعلق به یک خوشه نسبت به نمونه های متعلق به خوشه های دیگر تشابه بیشتری دارند.

تحلیل خوشه ای: ابزاری برای اکتشاف ساختار داده هاست بدون فرضیاتی که در روش های آماری در نظر می گیریم.

الگوریتم های خوشه بندی ابزارهای هستند که در جریان پیدا کردن ساختار داده ها مورد استفاده قرار می گیرند.

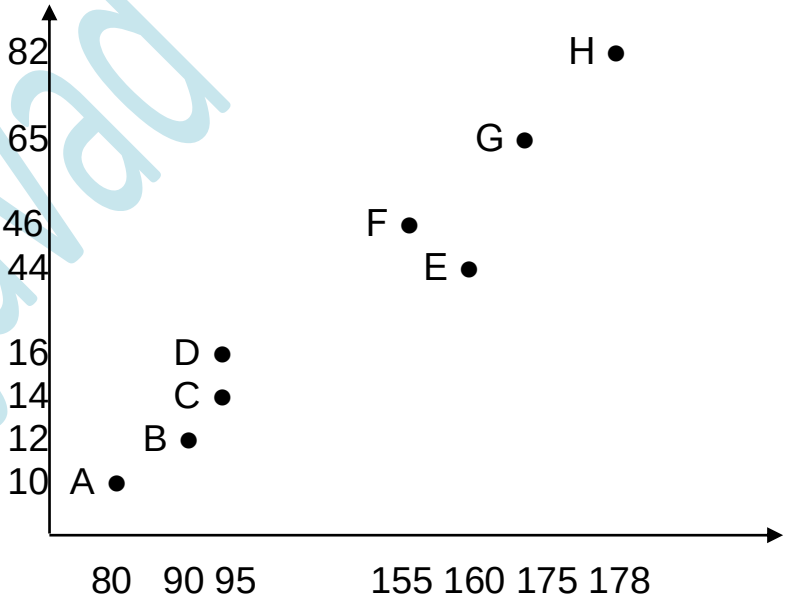
تحلیل خوشه ای روش پیدا کردن خوشه ها در داده ها نیز نامیده می شود.

مثال: مقادیر قد و وزن ۸ نفر در جدول (۱) داده شده است و می خواهیم این ۸ نفر را از نظر تشابهات وزنی و قدی به گروه هایی افراز کنیم.

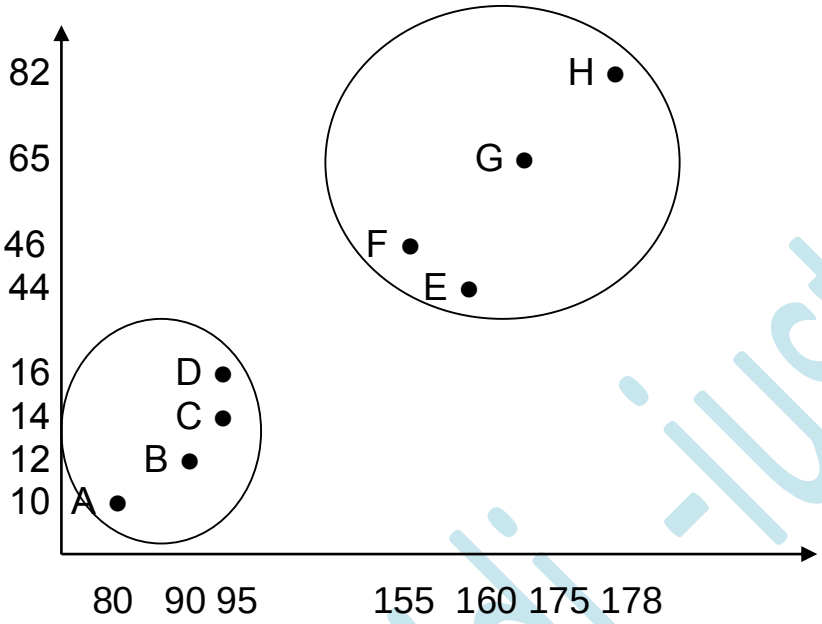
جدول(۱) : مقادیر قد و وزن ۸ نفر

	صفت		
		قد(cm)	وزن(kg)
افراد	A	۸۰	۱۰
	B	۹۰	۱۲
	C	۹۵	۱۴
	D	۹۵	۱۶
	E	۱۶۰	۴۴
	F	۱۵۵	۴۶
	G	۱۷۵	۶۵
	H	۱۷۸	۸۲

چون اشخاص را برای دو صفت مورد بررسی قرار داده ایم، می توانیم آنها را در یک نمودار دو بعدی نمایش دهیم.
با توجه به نمودار زیر وجود دو گروه بوضوح مشخص است، (A,B,C,D) در یک گروه و(E,F,G,H) در گروه و بزرگسالان در گروه دیگر قرار گرفته اند.
چنین گروه هایی را خوشه می نامند و کشف چنین گروه هایی هدف تحلیل خوشه ای می باشد.
اساساً، پژوهشگران می خواهند خوشه هایی را تشکیل دهند که موضوعات در خوشه ها نسبت به یکدیگر تشابه دارند، در صورتی که موضوعات در خوشه های مختلف(دو خوشه نسبت به یکدیگر) نامتشابه می باشند.



نمودار(۱):نمودار جدول (۱) بدون خوشه بندی



نمودار(۱):نمودار جدول (۱) در دو خوشه

همانطور که قبلاً گفتیم خوشه بندی زیر مجموعه ای از دسته بندی می باشد که به آن یادگیری بدون ناظر نیز می گویند و تعداد خوشه ها در آن مشخص نیست. به عنوان مثال فرض کنید شاخص های مختلفی از سلامتی فردی، افراد سیگاری و غیر سیگاری جمع آوری شده است. یک رده بندی بدون ناظر این افراد را بر مبنای تشابهات بین شاخص های سلامتی گروه بندی می نماید و سپس سعی می کند که تعیین کند آیا سیگاری بودن عاملی در سوق دادن افراد به سمت بیماری های مختلف است یا خیر. رده بندی با نظارت راه های تشخیص سیگاری ها از غیر سیگاری ها را بر اساس شاخص های سلامتی مطالعه می کند.

در واقع رده بندی بدون ناظر جوهره تحلیل خوشه ای است.

طبقه بندی موضوعات متشابه به گروه ها یک فعالیت مهم است، در زندگی روزمره، این امر قسمتی از فرآیند یادگیری می باشد: یک کودک تشخیص بین گربه و سگ، صندلی و میز، زن و مرد را یاد می گیرد و بدین وسیله رویه طبقه بندی ناخودآگاه به طور پیوسته بهبود می یابد(با این توضیح روشن می شود که چرا تحلیل خوشه ای به عنوان شاخه ای از تشخیص الگو و هوش مصنوعی مطرح می شود).

طبقه بندی همواره نقش حساسی در علوم داشته است. در دهه ۸۰ لینایوس و ساوجس طبقه بندی گسترده ای از حیوانات، گیاهان، معادن و بیماری ها را تهیه کردند. در علم نجوم هرتزپرنگ و راسل ستاره ها را بر اساس دو متغیر: فزونی نورشان و دمای سطح، آنها را در گروه های مختلف طبقه بندی کردند. در علوم اجتماعی، خیلی اوقات مردم بر اساس اولویت ها و رفتارشان طبقه بندی می شوند. اغلب در بازاریابی، بخش های بازار مشخص می شود(گروه هایی از مشتریان با نیازهای متشابه تشکیل می شوند). مثال های بیشتری در زمینه های علومی مانند

جغرافیا(خوشه بندی نواحی)، پزشکی(انواع سرطان ها)، شیمی(رده بندی ترکیبات)، باستان شناسی(گروه بندی یافته های باستان شناسی) و غیره، می توان ارائه داد.

توجه کنید که تحلیل خوشه ای کاملاً متفاوت از تحلیل ممیزی می باشد، بدین طریق که در تحلیل خوشه ای دسته ها را تشکیل می دهیم، در صورتی که در تحلیل ممیزی گروه ها از پیش تعیین شده اند و موضوعات را به گروه ها نسبت می دهیم.

الگوریتم ها و برنامه های کامپیوتری بسیار زیادی برای تحلیل خوشه ای وجود دارند که دلیل تعدد آنها احتمالاً در دو مورد زیر است:

نخست: تحلیل خوشه ای از سرعت رشد زیادی برخوردار بوده و در بین علوم مختلف گسترده شده است، که این امر را از هزاران مقاله چاپ شده در مجلات علمی می توان فهمید و این باعث شده که پژوهشگران علوم از رشد آن در علوم دیگر بی خبرند، برای همین اطلاعات دقیقی از این که توسط چه کسانی توسعه داده شده و یا در کدام یک از علوم رشد کرده، موجود نیست.

دوم: تعریف کلی برای خوشه وجود ندارد، و در حقیقت چندین نوع خوشه وجود دارد:

خوشه کروی، خوشه خطی، خوشه محدب و غیره

مجموعه گسترده ای از روش ها و همچنین نرم افزارهای مختلف موجود، به راحتی ذهن کاربر را مشغول این نکته می کند که یک دیدگاه مناسب برای حل مساله اش انتخاب کند، ولی معیاری که نشان دهد تکنیکی بر تکنیک دیگر برتری دارد موجود نیست و کاربرها از روی مسائل و نوع داده بهترین روش را انتخاب می کنند.

اگر تحلیل خوشه ای قادر به تشخیص ساختار موجود در داده ها نباشد، در عوض می تواند ساختاری را بوجود آورد که مجموعه داده ها از هم جدا شده با یک روش نسبتاً مناسب کم و بیش همگن شوند.

پیش نیازها

اصولی که تحلیل خوشه ای باید در آنها صدق کند عبارتند از:

۱. مقیاس پذیری.
۲. مناسب بودن برای انواع مختلف از صفات.
۳. کشف خوشه هایی با شکل های مختلف.
۴. پیش نیازهای مینیمال برای شناخت زمینه برای تعیین پارامترهای ورودی.
۵. توانایی برخورد با داده های دور افتاده و پرت.
۶. حساس نبودن به ترتیب داده های ورودی.
۷. قابلیت تفسیر و کاربرد.

مشکلات

۱. تکنیک های خوشه بندی به اندازه کافی پاسخگوی همه نیازها نیست.
۲. بررسی تعداد زیاد از داده ها با بعد زیاد به خاطر پیچیدگی زمانی مشکل آفرین است.
۳. موثر بودن روش، به فاصله تعریف شده بستگی دارد.

۴. اگر یک اندازه فاصله واضحی وجود نداشته باشد، باید آن را تعریف کرد. که تعریف آن، به خصوص در فضای چند بعدی کار آسانی نیست.
۵. نتایج الگوریتم های خوشه بندی می تواند، به صورت های مختلف قابل تفسیر باشد.

نکته: الگوریتم های خوشه بندی، نمونه ها یا داده ها را براساس اندازه مشابهت بین دو نمونه گروه بندی می کنند. مجموعه نمونه ها شامل داده های خام هستند که برای تحلیل خوشه ای ما نیاز به دو فرم استاندارد : ماتریس الگو و ماتریس نزدیکی داریم.

ماتریس الگو

نمونه ها معمولا بوسیله بردارهایی (نقاطی) در یک فضای چند بعدی نمایش داده می شوند، که هر بعد یک صفت (متغییر) مشخص از نمونه ها را نشان می دهد، مانند وزن، طول، جنسیت، رنگ و غیره.

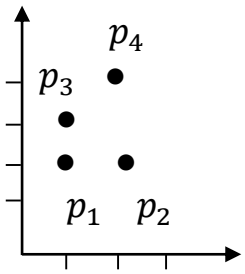
اگر فرض کنیم n نمونه داشته باشیم که هر کدام دارای d صفت باشند، پس یک ماتریس نمونه $n \times d$ را می توانیم تشکیل دهیم که هر سطر این ماتریس یک نمونه و هر ستون یک صفت مربوط به نمونه ها می باشد. نام دیگر این ماتریس، ماتریس داده می باشد. تاکر (۱۹۶۴) عبارت ماتریس دو وجهی را بکار برده است (چون درایه های سطرها و ستون ها متفاوت اند).

$$\begin{bmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{bmatrix}$$

مثال : اگر نمونه ها تراشه های تولید شده توسط یک دستگاه باشند دو صفت مورد مطالعه می تواند قطر و طول تراشه ها باشد که جدول (۲) ماتریس نمونه آن را نشان می دهد.

جدول (۲) : ماتریس الگوی متناظر چهار نمونه

	صفت		
		X	Y
تراشه	p_1	۱	۲
	p_2	۲	۲
	p_3	۱	۳
	p_4	۲	۴



نمودار(۲) : نمودار جدول (۲)

ماتریس نزدیکی

در روش های خوشه بندی برای ادغام کردن دو خوشه به ماتریسی به نام ماتریس نزدیکی نیازمندیم. این ماتریس که یک ماتریس یک وجهی $n\times n$ می باشد (چون سطر ها با ستون های متناظر یکی هستند)، با استفاده از ماتریس نمونه ساخته می شود و مؤلفه سطر iام و ستون jام آن، p_{ij} یا: اندازه مشابهت (دو نمونه همانند یکدیگرند مانند ضریب همبستگی) هستند یا اندازه عدم مشابهت (دو نمونه چقدر از یکدیگر متفاوتند مانند فاصله اقلیدسی) می باشند. به وضوح این ماتریس متقارن است.

مثال : برای ماتریس الگو جدول(۲) اگر اندازه نزدیکی را فاصله $p_{ij}=\sqrt{(x_i-x_j)^2+(y_i-y_j)^2}$ در نظر بگیریم، ماتریس نزدیکی به صورت زیر در می آید:

$$\begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array}\begin{array}{c} \begin{array}{cccc} p_1 & p_2 & p_3 & p_4 \end{array} \\ \begin{bmatrix} 0 & \sqrt{2} & 1 & \sqrt{5} \\ \sqrt{2} & 0 & 1 & 1 \\ 1 & 1 & 0 & \sqrt{2} \\ \sqrt{5} & 1 & \sqrt{2} & 0 \end{bmatrix} \end{array}$$

اگر اندازه نزدیکی را فاصله اقلیدسی در نظر بگیریم یک اندازه عدم مشابهت داریم، بدین صورت که هر چه فاصله اقلیدسی بین دو نمونه بیشتر باشد عدم مشابهت بین دو نمونه بیشتر می شود و هر چه فاصله کمتر شود عدم مشابهت نیز کمتر و برعکس آن اندازه مشابهت دو نمونه بیشتر می شود.

ماتریس نزدیکی برای داده های کمی مناسب بوده و داده های کیفی را در برنمی گیرد، در حالی که ماتریس مشابهت فراگیر است. لذا برای موارد آینده ذکر ماتریس مشابهت لازم به نظر می رسد.

برای اندازه عدم مشابهت توابع فاصله فراوانی مورد استفاده قرار می گیرد(جدول (۳)). هر یک از این فاصله ها، شکل هندسی مربوط به خود را دارد. تابع فاصله انتخاب شده، ویژگی های هندسی ویژه ای را برای خوشه ها به همراه دارد.

جدول(۳) : توابع فاصله میان گروه های X وY

تابع فاصله	فرمول و توضیحات
فاصله اقلیدسی	$d(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$
فاصله همینگ	$d(x,y)=\sum_{i=1}^{n}\lvert x_i-y_i\rvert$
فاصله چبیشف	$d(x,y)=\max_{i=1,2,\dots,n}\lvert x_i-y_i\rvert$
فاصله مینکوفسکی	$d(x,y)=\sqrt[p]{\sum_{i=1}^{n}(x_i-y_i)^p}\ p>0$
فاصله کانبرا	$d(x,y)=\sum_{i=1}^{n}\frac{\lvert x_i-y_i\rvert}{x_i+y_i}$ x_i و y_i مثبت می باشند
جدایی زاویه ای	$d(x,y)=\frac{\sum_{i=1}^{n}x_iy_i}{[\sum_{i=1}^{n}x_i^2\sum_{i=1}^{n}y_i^2]^{\frac{1}{2}}}$
تذکر: زاویه بین بردارهای یکه ای که در جهت x و y هستند، مد نظر است.	

برای اندازه مشابهت از ضریب همبستگی استفاده می شود.اگر صفت برای نمونه ها به صورت دو حالتی باشد، چندین نوع اندازه مشابهت می توان تعریف کرد. جدول (۴) را برای اندازه مشابهت تشکیل می دهیم:

	x_k		
		1	0
x_i	1	a	b
	2	c	d

جدول (۴)

a= تعدادی از صفات که برای هر دو نمونه یک می باشند.

b= تعدادی از صفات که برای نمونه iام صفر و برای نمونه kام یک هستند.

c= تعدادی از صفات که برای نمونه iام یک و برای نمونه kام صفر هستند.

d= تعدادی از صفات که برای هر دو نمونه صفر می باشند.

ساده ترین معیار سنجش، عبارت است از ضریب تطابق که به صورت زیر تعریف می شود

$$\frac{a+d}{a+b+c+d}$$

یکی دیگر از معیار های سنجش تشابه را راسل و رائو تعریف کرده اند:

$$\frac{a}{a+b+c+d}$$

در شاخص جاکارد مواردی در نظر گرفته می شود که در آنها، حداقل یکی از دو ورودی مقدار ۱ را اختیار کرده است:

$$\frac{a}{a+b+c}$$

شاخص چکانوفسکی تا حدودی مشابه با شاخص قبلی می باشد. اما، این شاخص با اضافه کردن ضریب وزنی ۲، رخ دادن وضعیت هایی را مورد تاکید قرار داده است که در آنها هر دو درایه موجود در x و y مقدار ۱ را اختیار می کنند:

$$\frac{2a}{2a+b+c}$$

در رابطه با داده های دودویی، استفاده از فاصله همینگ نیز می تواند مفید واقع شود. آگاهی از رویکردهای متعدد در تعریف معیارهای سنجش تشابه از اهمیت فراوانی برخوردار است. در ضمن، انتخاب معیار سنجش تشابه به نوع کاربرد مورد نظر نیز بستگی دارد.

مثال : یک آزمون ۲۰ سوالی که جواب های آن بلی (۱) یا خیر (۰) است برای دو نفر برگزار گردیده است، که نتیجه آن در جدول زیر است.

	سوال																					
نفر		۱	۲	۳	۴	۵	۶	۷	۸	۹	۱۰	۱۱	۱۲	۱۳	۱۴	۱۵	۱۶	۱۷	۱۸	۱۹	۲۰
	x_1	۰	۱	۱	۰	۰	۱	۰	۰	۱	۰	۰	۱	۱	۱	۰	۰	۱	۰	۱	۰
	x_2	۰	۱	۱	۱	۰	۰	۰	۰	۱	۱	۱	۱	۱	۱	۰	۱	۱	۰	۱	۰

برای محاسبه اندازه مشابهت و عدم مشابهت جدول زیر را تشکیل می دهیم

	x_2		
		1	0
x_1	1	۸	۱
	۰	۴	2۷

در نتیجه داریم

	مشابهت	عدم مشابهت
ضریب تطابق	$\frac{15}{20}=0/75$	$\frac{5}{20}=0/25$
ضریب جاکارد	$\frac{8}{13}=0/615$	$\frac{5}{13}=0/385$

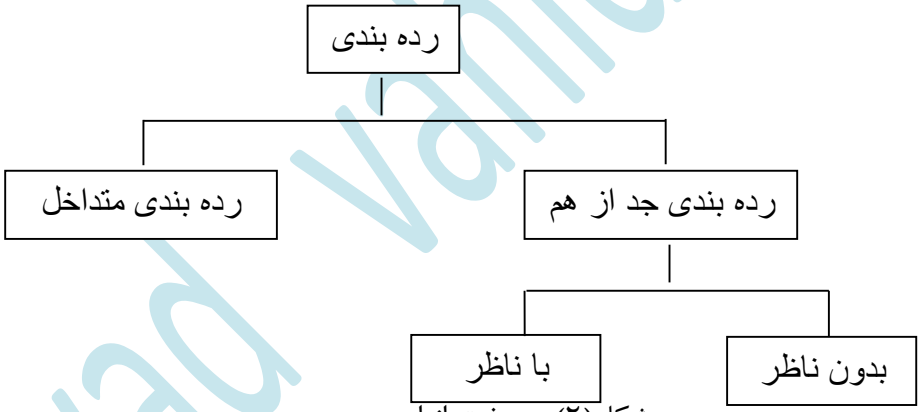
جدول(۵)

انواع رده بندی

تحلیل خوشه ای فرآیند رده بندی نمونه ها به زیر مجموعه هایی است که در رابطه با مسئله خاصی دارای معنی و مفهوم می باشند. این نمونه ها به طور مؤثری سازماندهی می شوند تا جامعه ای را که نمونه گیری می کنیم مشخص نمایند. الگوریتم خوشه بندی سلسله مراتبی عبارت از مشخص کردن گام هایی برای اجرای خوشه بندی سلسله مراتبی است. اغلب بهتر است که روش خوشه بندی سلسله مراتبی را با برقراری یک الگوریتم مشخص کنیم. ولی باید الگوریتم از خود روش جدا شده باشد. در این قسمت علاوه بر تعریف الگوریتم ها و روش ها ، نوع ساختار ریاضی را که یک خوشه بندی سلسله مراتبی بر داده ها تحمیل می کند تعریف کرده و راه های نگاه به این مسئله را توضیح می دهیم.

خوشه بندی یک نوع خاص رده بندی روی یک مجموعه متناهی از نمونه ها می باشد. شکل زیر یک درخت از مسائل رده بندی را نشان می دهد.

هر برگ این درخت یک نوع مختلف از مسائل رده بندی را مطرح می کند. گره های درخت به صورت زیر تعریف می شوند.



شکل(۲) : درخت انواع رده بندی

رده بندی جدا از هم و رده بندی متداخل

رده بندی جدا از هم افرازی از مجموعه نمونه ها است که هر نمونه می تواند فقط متعلق به یک زیر مجموعه یا خوشه باشد، ولی در رده بندی متداخل یک نمونه می تواند در چندین زیر مجموعه قرار بگیرد. برای مثال اگر یک گروه از افراد را بخواهیم نسبت به صفات سن و جنس خوشه بندی کنیم خوشه بندی جدا از هم است، در صورتی که اگر این افراد را نسبت به نوع بیماری خوشه بندی کنیم خوشه بندی متداخل است، زیرا ممکن است یک فرد همزمان به چندین بیماری مبتلا باشد.

یک نوع از خوشه بندی متداخل، خوشه بندی فازی است که در آن هر نمونه می تواند با یک درجه عضویت متعلق به چند خوشه باشد. در اینجا ما فقط با رده بندی غیر فازی روبرو هستیم.

رده بندی بدون ناظر و با ناظر

رده بندی بدون ناظر برای انجام رده بندی فقط از ماتریس مشابهت استفاده می کند. این نوع رده بندی را در تشخیص الگو، یادگیری بدون ناظر می گویند، زیرا هیچ برچسب رده ای که یک افراز پیشاپیش از نمونه ها را نمایش دهد مورد استفاده قرار نمی دهد، ولی رده بندی با ناظر برچسب های مربوط به نمونه ها و همین طور ماتریس مشابهت را مورد استفاده قرار می دهد. مبحث ما رده بندی بدون ناظر جدا از هم می باشد که به آن خوشه بندی می گویند.

نکته: خوشه بندی در واقع یک عملیات غیر نظارتی می باشد. این عملیات هنگامی استفاده می شود که ما به دنبال یافتن گروه هایی از داده های مشابه می باشیم بدون اینکه از قبل یک پیش بینی در مورد شباهت های موجود داشته باشیم . هر خوشه شامل مجموعه ای از اشیاء داده ای است که به هم شبیه هستند اما با اشیاء خارج از آن متفاوت می باشند .

برای اینکه بتوانیم داده ها را خوشه بندی کنیم باید بتوانیم میزان شباهت آنها را بدست آوریم. اینکار معمولا به صورت غیر مستقیم انجام می شود، یعنی با استفاده از معیارهای اندازه گیری فاصله، میزان تفاوت بین دو شی بدست می آید بطوریکه اشیاء شبیه به هم دارای فاصله کمتری خواهند بود. چندین معیار اندازه گیری فاصله برای خوشه بندی وجود دارد که معروفترین آن فاصله اقلیدسی است.

طبقه بندی الگوریتم های خوشه بندی

می توان تکنیک های خوشه بندی را به دو گروه اصلی تقسیم کرد:

۱- خوشه بندی سلسله مراتبی

۲- خوشه بندی افرازی

خوشه بندی سلسله مراتبی

الگوریتم های خوشه بندی سلسله مراتبی، داده ها را به صورت یک درخت نمایش می دهد که به این درخت سلسله مراتبی دندروگرام می گویند. دندروگرام مرکب از لایه هایی از گره هاست که هر کدام یک خوشه را نمایش می دهد.

ساخت این درخت با توجه به دو رویکرد انجام می شود:پایین به بالا و بالا به پایین

در رویکرد پایین به بالا که به آن رویکرد انباشتی یا تجمعی نیز گفته می شود، هر داده‌ها به عنوان خوشه‌ای مجزا در نظر گرفته می‌شود و در طی فرایندی تکراری در هر مرحله خوشه‌هایی که شباهت بیشتری با یکدیگر با یکدیگر ترکیب می‌شوند تا در نهایت یک خوشه و یا تعداد مشخصی خوشه حاصل شود.

در رویکرد بالا به پایین که به آن رویکرد تقسیم کننده نیز گفته می شود، تمام داده‌ها به عنوان یک خوشه در نظر گرفته می‌شوند و سپس در طی یک فرایند تکراری در هر مرحله داده‌هایی شباهت کمتری به هم دارند به خوشه‌های مجزایی شکسته می‌شوند و این روال تا رسیدن به خوشه‌هایی که دارای یک عضو هستند ادامه پیدا می‌کند.
روش های خوشه بندی سلسله مراتبی تجمعی از الگوریتم کلی زیر استفاده می کنند که در آن $P=[p(i,j)]$ یک ماتریس مشابهت $n\times n$ و $L(k)$ ، $k=0,1,2,...,n-1$ سطح k امین خوشه بندی است.

الگوریتم جانسون برای خوشه بندی سلسله مراتبی تجمعی

گام ۱: با خوشه بندی جدا با سطح $L(0)=0$ و عدد دنباله ای $m=0$ آغاز کنید.

گام ۲: در خوشه بندی جاری دو خوشه دارای کمترین عدم مشابهت (بیشترین مشابهت) مثلاً جفت $\{(r),(s)\}$ را که در

$$p[(r),(s)] = \min\{p[(i),(j)]\}$$

صدق می کند، پیدا کنید که در آن مینیمم روی همه جفت از خوشه ها در خوشه بندی جاری می باشد.

گام ۳: شماره دنباله را افزایش دهید: $m+1 \leftarrow m$. برای تشکیل خوشه بندی بعدی m، خوشه های (r) و (s) را با هم ادغام کرده و یک خوشه تکی بوجود آورید. سطح این خوشه بندی را

$$L(m)=p[(r),(s)]$$

قرار دهید.

گام ۴: ماتریس مشابهت P را دوباره تشکیل دهید، بدین صورت که سطر و ستون متناظر باخوشه های (r) و (s) را حذف می کنیم و یک سطر و ستون برای خوشه ای که جدیداً تشکیل شده اضافه می کنیم. مشابهت بین خوشه جدید (r,s) و یکی از خوشه های قدیم (k) بصورت زیر تعریف می شود.

$$d[(k),(r,s)]=$$
$$\alpha_r d[(k),(r)]+\alpha_s d[(k),(s)]+\beta d[(r),(s)]+\gamma|d[(k),(r)]-d[(k),(s)]|$$

لنس و ویلیامز (۱۹۶۷) اولین بار این فرمول را پیشنهاد کردند. جدول (۶) مقادیر پارامترها، برای معمول ترین الگوریتم ها را نشان می دهد.

جدول (۶): مقادیر پارامترهای اندازه مشابهت الگوریتم سلسه مراتبی

		ضرایب			
		α_r	α_s	β	γ
روش خوشه بندی سلسله مراتبی	اتصال تک	$\frac{1}{2}$	$\frac{1}{2}$	۰	$-\frac{1}{2}$
	اتصال کامل	$\frac{1}{2}$	$\frac{1}{2}$	۰	$\frac{1}{2}$

گام ۵: اگر همه نمونه ها در یک خوشه باشند الگوریتم ها را متوقف کنید، در غیر اینصورت به گام ۲ بروید.

مثال: فرض می کنیم ماتریس نزدیکی براساس عدم مشابهت به شکل زیر می باشد.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{bmatrix} 0 & 2/3 & 3/4 & 1/2 & 3/7 \\ 2/3 & 0 & 2/6 & 1/8 & 4/6 \\ 3/4 & 2/6 & 0 & 4/2 & 0/7 \\ 1/2 & 1/8 & 4/2 & 0 & 4/4 \\ 3/7 & 4/6 & 0/7 & 4/4 & 0 \end{bmatrix} \end{array}$$

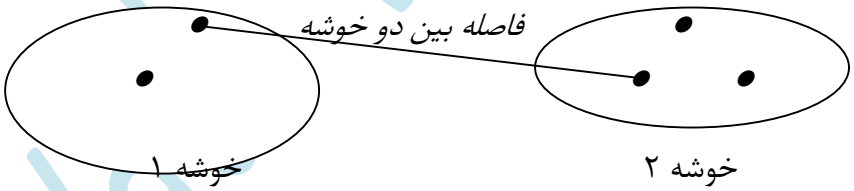
در نتیجه خوشه ها را بر اساس مینیمم با هم ادغام می کنیم ولی اگر ماتریس براساس اندازه مشابهت باشد، خوشه ها را براساس ماکزیمم با هم ادغام می کنیم، چون دو خوشه ای با هم ادغام می شوند که دارای بیشترین مشابهت باشند.

خوشه بندی سلسله مراتبی تجمعی اتصال تک

مشابهت بین خوشه جدید (r,s) و خوشه قدیم (k) براساس پارامتر های جدول(۶) به صورت زیر می باشد

.

$$d[(k),(r,s)] = \frac{1}{2}d[(k),(r)] + \frac{1}{2}d[(k),(s)] - \frac{1}{2}|d[(k),(r)] - d[(k),(s)]|$$
$$= \min\{d[(k),(r)], d[(k),(s)]\}$$



شکل (۳) : فاصله بین دو خوشه در روش اتصال تک مینیمم فاصله ها می باشد

حال طبق مراحل الگوریتم داریم $\min p_{ij} = p_{35} = 0.7$ در نتیجه خوشه (۵) و (۳) را به عنوان یک خوشه جدید در نظر می گیریم پس برای تشکیل ماتریس مشابهت چهار خوشه (۱)،(۲)،(۵,۳)و(۴) را داریم، فاصله خوشه جدید از دیگر خوشه ها را به شکل زیر بدست می آوریم.

$$P_{(3,5)1} = \min(P_{13}, P_{15}) = \min(3.4, 3.7) = 3.4$$
$$P_{(3,5)2} = \min(P_{23}, P_{25}) = \min(2.6, 4.6) = 2.6$$
$$P_{(3,5)4} = \min(P_{43}, P_{45}) = \min(4.2, 4.4) = 4.2$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3,5 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3,5 & 4 \end{array} \\ \begin{bmatrix} 0 & 2/3 & 3/4 & 1/2 \\ 2/3 & 0 & 2/6 & 1/8 \\ 3/4 & 2/6 & 0 & 4/2 \\ 1/2 & 1/8 & 4/2 & 0 \end{bmatrix} \end{array} \Longrightarrow \min p_{ij} = 1/2$$

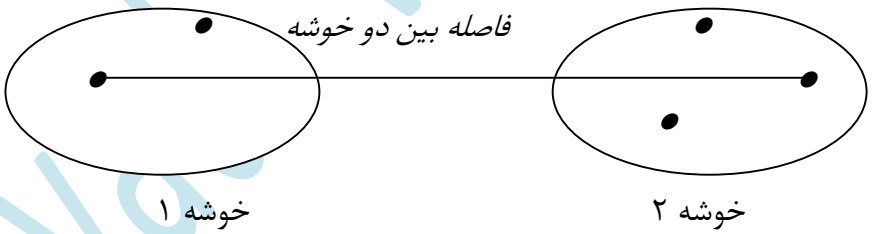
$$\begin{array}{c} \\ 1,4 \\ 2 \\ 3,5 \end{array}\begin{array}{c} \begin{array}{ccc} 1,4 & 2 & 3,5 \end{array} \\ \begin{bmatrix} 0 & 1/8 & 3/4 \\ 1/8 & 0 & 2/6 \\ 3/4 & 2/6 & 0 \end{bmatrix} \end{array} \Longrightarrow \min p_{ij} = 1/8$$

$$\begin{array}{c} \\ 1,4,2 \\ 3,5 \end{array}\begin{array}{c} \begin{array}{cc} 1,4,2 & 3,5 \end{array} \\ \begin{bmatrix} 0 & 2/6 \\ 2/6 & 0 \end{bmatrix} \end{array}$$

خوشه بندی سلسله مراتبی تجمعی اتصال کامل

عدم مشابهت بین خوشه جدید (r,s) و خوشه قدیم (k) براساس پارامترهای جدول(۶) به صورت زیر است.

$$d[(k),(r,s)] = \max\{d[(k),(r)], d[(k),(s)]\}$$



شکل (۴) : فاصله بین دو خوشه در روش اتصال کامل ماکسیمم فاصله ها می باشد

حال طبق مراحل الگوریتم داریم $\min p_{ij} = p_{35} = 0.7$ در نتیجه خوشه (۵) و (۳) را به عنوان یک خوشه جدید در نظر می گیریم برای تشکیل ماتریس مشابهت چهار خوشه (۱)،(۲)،(۳،۵)و(۴) را داریم، فاصله خوشه جدید از دیگر خوشه ها را به شکل زیر به دست می آوریم.

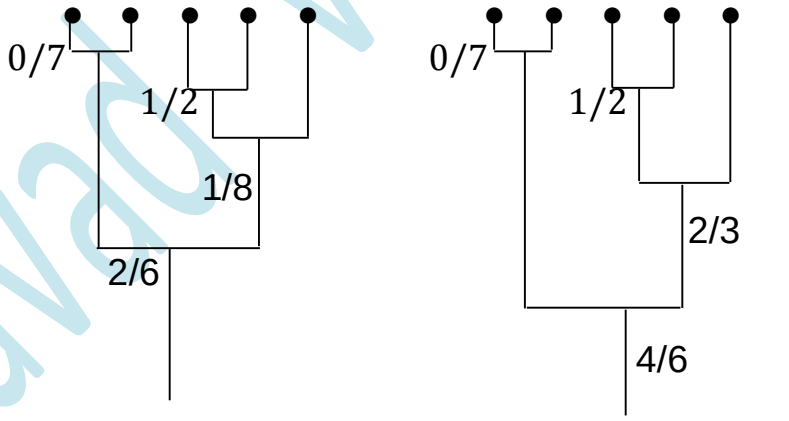
$$P_{(3,5)1} = \max(P_{13}, P_{15}) = \max(3.4, 3.7) = 3.7$$
$$P_{(3,5)2} = \max(P_{23}, P_{25}) = \max(2.6, 4.6) = 4.6$$
$$P_{(3,5)4} = \max(P_{34}, P_{45}) = \max(4.2, 4.4) = 4.4$$

$$\begin{array}{c} \\ 1 \\ 2 \\ 3,5 \\ 4 \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3,5 & 4 \end{array} \\ \begin{bmatrix} 0 & 2/3 & 3/7 & 1/2 \\ 2/3 & 0 & 4/6 & 1/8 \\ 3/7 & 4/6 & 0 & 4/4 \\ 1/2 & 1/8 & 4/4 & 0 \end{bmatrix} \end{array} \Longrightarrow \min p_{ij} = 1/2$$

$$\begin{array}{c} \\ 1,4 \\ 2 \\ 3,5 \end{array}\begin{array}{c} \begin{array}{ccc} 1,4 & 2 & 3,5 \end{array} \\ \begin{bmatrix} 0 & 2/3 & 4/4 \\ 2/3 & 0 & 4/6 \\ 4/4 & 4/6 & 0 \end{bmatrix} \end{array} \Longrightarrow \min p_{ij} = 1/8$$

$$\begin{array}{c} \\ 1,4,2 \\ 3,5 \end{array}\begin{array}{c} \begin{array}{cc} 1,4,2 & 3,5 \end{array} \\ \begin{bmatrix} 0 & 4/6 \\ 4/6 & 0 \end{bmatrix} \end{array}$$

دندروگرام اتصال تک و کامل به صورت زیر می باشد.



دندروگرام اتصال تک　　دندروگرام اتصال کامل

نمودار(۳) : دندروگرام اتصال تک و اتصال کامل

خوشه بندی افرازی

تکنیک های خوشه بندی سلسله مراتبی داده ها را به صورت دنباله ای لانه ای از گروه ها سازمان دهی می کند. مشخصه مهمی از روش های خوشه بندی سلسله مراتبی تاثیر بصری دندروگرام است که به تحلیل گر داده ها این امکان را می دهد که مشاهده کند چگونه نمونه ها در سطوح متوالی مشابهت در خوشه ها ادغام یا تفکیک می شوند.

نوع دیگر از روش خوشه بندی، خوشه بندی غیر سلسله مراتبی یا روش افرازی می باشد. در این روش خوشه بندی، یک افراز تک از داده ها ایجاد می شود. هر دو روش سلسله مراتبی و افرازی زمینه های کاربرد مخصوص به خود را دارند، روش های خوشه بندی سلسله مراتبی فقط نیاز به ماتریس مشابهت دارند، در صورتی که روش های افرازی علاوه بر آن به ماتریس الگو نیز نیاز دارند.

روش های خوشه بندی سلسله مراتبی بیشتر در علوم زیست شناسی، اجتماعی، رفتاری به دلیل نیاز به ساختن رده بندی های بسیار، بیشتر طرفدار دارد، و تکنیک های افرازی در کاربردهای مهندسی که در آنها افرازهای تکی اهمیت دارند مورد استفاده واقع می شود.

روش های افرازی خوشه بندی بویژه برای نمایش و فشردگی موثر پایگاه داده های حجیم بسیار مناسب است. دندروگرام ها برای بیشتر از چند صد نمونه غیر عملی است.

مسئله خوشه بندی افزاری را می توان به صورت کلی زیر بیان کرد:

n نمونه در یک فضای متریک d بعدی مفروض اند، افرازی از نمونه ها را به صورت k گروه یا خوشه تعیین کنید به طوری که نمونه های متعلق به یک خوشه بیشتر از نمونه های خوشه های مختلف دیگر به یکدیگر شبیه باشند. مقدار k تعداد خوشه ها، می تواند از قبل مشخص باشد یا نباشد.

یک معیار خوشه بندی، مانند مربعات خطا، باید انتخاب شود. معیارها را می توان به صورت فراموضعی و موضعی طبقه بندی کرد. یک معیار فراموضعی در ابتدا k نمونه را انتخاب می کند و خوشه ها را بدین شکل تشکیل می دهد که هر نمونه ای که به یکی از این k نمونه اولیه متشابه تر باشد در آن خوشه قرار می گیرد. ولی معیار موضعی، k خوشه را بر حسب تراکم ناحیه ها تعیین می کند یا k ناحیه ای که دارای نزدیکترین همسایگی ها هستند.

راه حل نظری این مساله افرازی مستقیم است. یک معیار را انتخاب کرده و آن را به ازای کلیه افرازهای ممکن شامل k خوشه ارزیابی کنید و افرازی را که این معیار را بهینه می سازد انتخاب نمائید.

اولین مشکلی که خوشه بندی افرازی با آن مواجه می شود انتخاب یک معیار است که مفاهیم شهودی فرد درباره خوشه را، بصورت فرمول های ریاضی معقول ترجمه کند. معیارها شدیداً به پارامترهای مسئله بستگی دارد و به دلایل محاسباتی باید ساده، ولی به اندازه کافی پیچیده باشند تا ساختارهای داده های گوناگون را بازتاب دهند.

دومین مشکل این روش ارقام نجومی این افرازها حتی برای تعداد متوسط داده ها می باشد، که بدین ترتیب ارزیابی حتی ساده ترین معیار نسبت به کلیه افرازها غیر عملی است.

فرض کنید $S(n,k)$ تعداد خوشـه بندی های n نمونه بصورت k خوشه را نشان می دهد. از ترتیب نمونه ها در هر خوشه و ترتیب خود خوشه ها و تعداد خوشه های تهی صرف نظر شده است. یک معادله تفاضلی جزیی را می توان به صورت زیر برای $S(n,k)$ نوشت. فرض کنید همه خوشه بندی های متشکل از n-1 نمونه فهرست شده اند. یک خوشه بندی مرکب از n نمونه را، از این فهرست می توان به دو طریق تشکیل داد.

۱. نمونه nام را می توان بعنوان یک خوشه تک به هر عضو این لیست با دقیقاً k-1 خوشه اضافه کرد.

۲. نمونه nام را می توان به هر خوشه عضو دلخواه این فهرست با دقیقاً k خوشه اضافه کرد.

بنابراین شرایط مرزی بر روی معادله عبارتند از:

$$S(n,K)=S(n-1,K-1)+KS(n-1,K)$$

$$s.t. \quad S(n,1)=1, S(n,n)=1, S(n,K)=0 \quad if \quad K>n$$

برای حل معادله $S(n,k)$ نیاز است مقادیر $\{S(j,p)\}$ برای

$$\{(j,p): \quad 1\leq j\leq n-2 \quad 1\leq p\leq K\}$$

معلوم باشند.

جواب معادله تفاضلی جزیی(اعداد استرلینگ از نوع دوم)به صورت زیر است:

$$S(n,K)=\frac{1}{K!}\sum_{i=1}^{K}(-1)^n\binom{K}{i}(i)^n$$

همانطور که محاسبات نشان می دهند برای افراز ۱۰ نمونه به ۴ خوشه ۳۴۱۰۵ افراز داریم و این عدد برای افراز ۱۹ نمونه به ۴ خوشه برابر است با ۱۱۲۵۹۶۶۶۰۰۰ که رشد خیلی زیاد داشته است.

پس تعیین همه افرازهای ممکن حتی برای تعداد کم از نمونه های کاری نا ممکن است.

یک راه برای دوری از این حجم محاسـبات، این است که یک تابع معیار، تنها برای مجموعه کوچکی از افرازهای قابل قبول ارزیابی می شود.

چگونه زیر مجموعه ای کوچک از افرازها را که دارای شـانس خوبی برای در بر داشـتن افراز بهینه می باشـند را مشـخص کنیم؟ معمول ترین رهیافت عبارت از بهینه کردن تابع معیار با استفاده از فن تکراری صعود از پله می باشد. با آغاز از یک افراز اولیه، نمونه ها را از خوشه ای به خوشه ای دیگر حرکت می دهیم با این هدف که مقدار تابع معیار را بهبود بخشیم. بنابراین هر افراز متوالی جابجایی از فراز قبلی است و بنابراین تنها یک تعداد کمی از فرازها بررسی می شوند.

راه دیگر دوری از انفجار ترکیباتی تا حدی مشـخص و رد کردن تعداد زیادی افرازهاست، که محتملاً مورد توجه نمی باشند(جنسن (۱۹۶۹) یک برنامه ریزی پویا برای این کار به کار برده است).

معیار خوشه بندی

به خاطر وجود نداشتن یک تعریف معین و عملی از خوشه، بهترین معیار تک برای بدست آوردن یک افراز وجود ندارد. از آن سو صدها توابع معیار پیشنهاد شده در ادبیات خوشه بندی مرتبت هستند و یک معیار یکسانی، به چندین صورت ظاهر می شود، و هر معیار خوشعه بندی ساختاری دیگر بر روی داده ها تحمیل می کند.

معیار خوشه بندی، مربعات خطا (K-means)

معیار مجموع مربعات خطا، K- میانگین در اکثر استراتژی های خوشه بندی بکار می رود. هدف کلی بدست آوردن آن افرازی است که مربعات خطا (اختلاف درون خوشه ها) را برای یک تعداد ثابت از خوشه ها مینیمم کند (یا ماکزیمم کردن اختلاف بین خوشه ها).

کلیات خوشه بندی افرازی

فرض کنید n نمونه d بعدی را به K خوشه $\{C_1, C_2, \dots, C_k\}$ به گونه ای افراز کنیم که C_k دارای n_k نمونه و هر نمونه فقط متعلق به یک خوشه است. به طوری که $\sum_{k=1}^{k} n_k = n$ و بردارمیانگین یا مرکز خوشه c_k به عنوان مرکزیت این خوشه تعریف می شود، یعنی

$$c_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} = \left(c_1^{(k)}, c_2^{(k)}, \dots, c_d^{(k)}\right)$$

که در آن $x_i^{(k)}$ نمونه iام متعلق به خوشه C_k است.

$$e_k^2 = \sum_{i=1}^{n_k} \|x_i^{(k)} - c_k\|^2$$

مربع خطای خوشه k ام است که برابر مجموع مربعات فاصله های اقلیدسی بین هر نمونه در C_k و مرکز خوشه اش c_k می باشد . این مربعات خطا تغییرات درون خوشه نیز نامیده میشود و

$$E_K^2 = \sum_{k=1}^{K} e_k^2$$

مجموع اختلاف بین خوشه هاست که باید مینیمم شود.

الگوریتم خوشه بندی افرازی تکراری (K- میانگین)

گام ۱: K خوشه دلخواه را به عنوان اولین افراز انتخاب کنید.

تکرار گام ۲ تا ۵ تا اینکه اعضای هر خوشه تثبیت شوند.

گام ۲: افراز جدید را با انتساب هر نمونه به نزدیکترین مرکز خوشه اش تشکیل دهید.

گام ۳: مرکز خوشه های جدید را محاسبه کنید.

گام ۴: مرحله ۲ و ۳ را تا بدست آوردن یک مقدار بهینه برای تابع معیار تکرار کنید.

گام ۵: تعداد خوشه ها را بوسیله یکی کردن و جدا کردن خوشه های موجود و یا بوسیله حذف خوشه های کوچک و دور افتاده تعدیل کنید.

مربعات خطا با افزایش تعداد خوشه ها کاهش پیدا می کندو باید برای تعداد ثابت از خوشه ها مینیمم شود.

اولین افراز (گام اول)

افراز اولیه می تواند بوسیله مجموعه ای از K نقطه هسته ای که بطور تصادفی از ماتریس الگوانتخاب شده اند تشکیل شود نقاط هسته ممکن است اولین K نمونه یا K نمونه ای که از ماتریس الگو به تصادف انتخاب شده اند باشند .مجموعه مرکب از K نمونه که کاملا از یکدیگر جدا باشند را می توان با انتخاب مرکزیت این داده ها به عنوان نقطه هسته اولیه و گزینش نقاط هسته متوالی که حداقل یک فاصله معینی از نقاط هسته قبلا انتخاب شده دارند به دست آورد .افراز اولیه یا خوشه بندی با انتساب اولیه هر نمونه به نزدیکترین نقطه هسته تشکیل می شود .

از خوشه بندی سلسله مراتبی نیز می توان برای تعیینی K خوشه استفاده کرد.

به روز کردن افراز (گام دوم)

افراز ها با انتساب مجدد نمونه ها به خوشه ها با تلاش برای کاهش مربع خطا به روز می شوند . برای این منظور باید مرکز خوشه ها محاسبه شده و نمونه به آن خوشه ای متعلق است که کمترین فاصله را از مرکز خوشه داشته باشد و این عمل تکرار می شود تا معیار همگرایی بدست آید و اعضای خوشه ها دیگر تغییر نکنند.

محاسبه مرکز خوشه های جدید (گام سوم)

چند روش برای محاسبه مرکز خوشه ها وجود دارد. مک کوئین روشی را ارائه کرد که در آن مرکز خوشه پس از اینکه نمونه ای به آن خوشه متعلق شد دوباره محاسبه شود ولی در روش فرجی مرکز خوشه پس از تعیین همه اعضای خوشه محاسبه می شود.

مثال : فرض کنید جدول زیر ماتریس الگوی ۱۰ مشاهده باشد برای تبدیل آن به K=2 خوشه ،

جدول (۷) : ماتریس الگو

شماره نمونه	صفت x	صفت y
۱	۱	۴
۲	۵	۱
۳	۵	۲
۴	۵	۴
۵	۱۰	۴
۶	۲۵	۴
۷	۲۵	۶
۸	۲۵	۷
۹	۲۵	۸
۱۰	۲۹	۷

اگر خوشه بندی افرازی باشد با توجه به جدول زیر (جدول فاصله نمونه ها از ۱و ۵) دو خوشه $C_1=\{1,2,3,4\}$ و $C_2=\{5,6,7,8,9,10\}$ به وجود می آید که میانگین فاصله ها ۹/۳۷ است.

جدول (۸) : انتساب نمونه ها به دو نمونه ۱و۵

شماره نمونه	عدم مشابهت از نمونه ۱	عدم مشابهت از نمونه ۵	کمترین عدم مشابهت	نزدیکترین نمونه
۱	۰	۹	۰	۱
۲	۵	۵/۸۳	۵	۱
۳	۴/۴۷	۵/۳۹	۴/۴۷	۱
۴	۴	۵	۴	۱
۵	۹	۰	۰	۵
۶	۲۴	۱۵	۱۵	۵
۷	۲۴/۰۸	۱۵/۱۳	۱۵/۱۳	۵
۸	۲۴/۱۹	۱۵/۱۳	۱۵/۱۳	۵
۹	۲۴/۳۳	۱۵/۵۲	۱۵/۵۲	۵
۱۰	۲۸/۱۶	۱۹/۲۴	۱۹/۲۴	۵

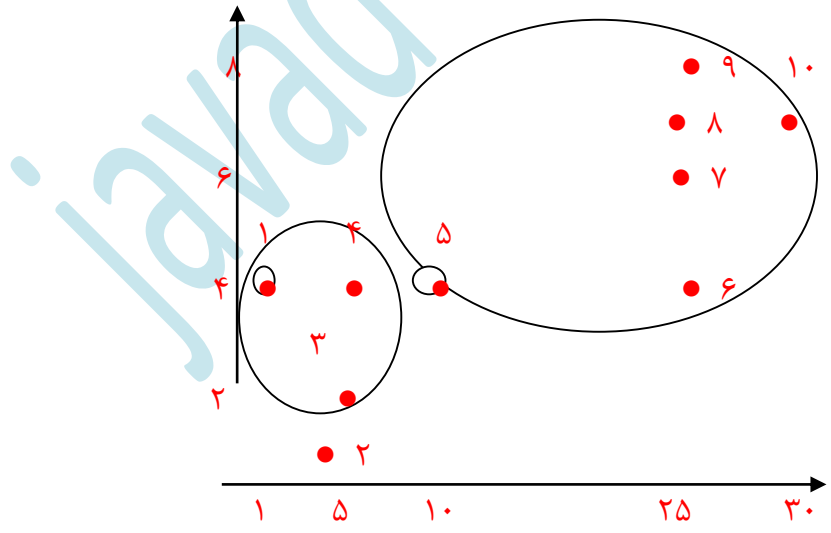
اگر با افراز ابتدایی ۸ و ۴ شروع شود جدول زیرو دو خوشه $C_1=\{1,2,3,4,5\}$ و $C_2=\{6,7,8,9,10\}$ با میانگین عدم مشابهت ۲/۳۰ بوجود می آید.

جدول (۹) : انتساب نمونه ها به دو نمونه ۴و۸

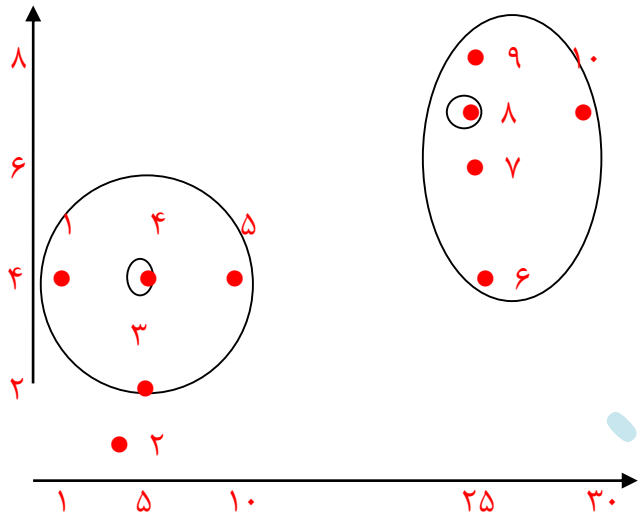
شماره نمونه	عدم مشابهت از نمونه ۱	عدم مشابهت از نمونه ۵	کمترین عدم مشابهت	نزدیکترین نمونه
۱	۴	۲۴/۱۹	۴	۴
۲	۳	۲۰/۸۸	۳	۴
۳	۲	۲۰/۶۲	۲	۴
۴	۰	۲۰/۲۲	۰	۴
۵	۵	۱۵/۳	۵	۴
۶	۲۰	۳	۳	۸
۷	۲۰/۱	۱	۱	۸
۸	۲۰/۲۲	۰	۰	۸
۹	۲۰/۴۰	۱	۱	۸
۱۰	۲۴/۱۹	۴	۴	۸

ولی اگر با افراز ابتدایی ۳ و ۸ الگوریتم خوشه بندی آغاز شود خوشه ها همان $C_1=\{1,2,3,4,5\}$ و $C_2=\{6,7,8,9,10\}$ می باشند، با این تفاوت که میانگین عدم مشابهت ها ۲/۱۹ بدست می آید.

نمودار زیر تصویری از دو افراز ابتدایی را نشان می دهد . به وضوح افراز ابتدایی که با ۴ و ۸ شروع شده بهتر و سریع تر همگراست.



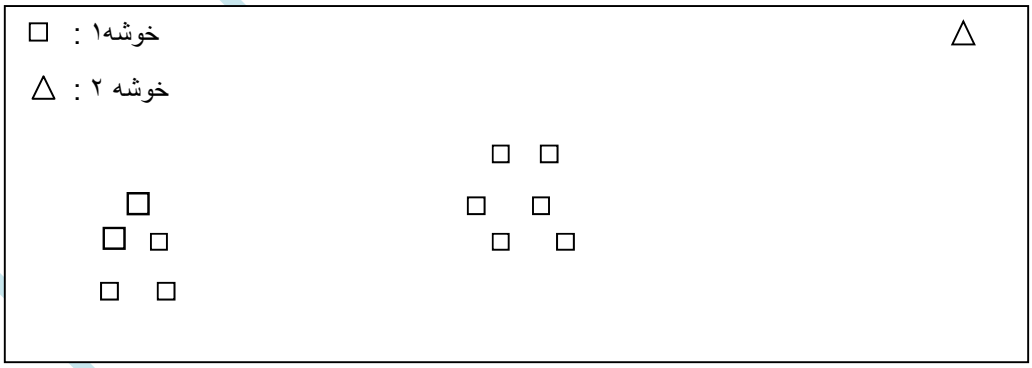
نمودار(۴) : خوشه بندی براساس جدول (۸)



نمودار(۴) : خوشه بندی براساس جدول (۹)

تعدیل تعداد خوشه ها (گام پنجم)

در بعضی مواقع پس از پایان یافتن خوشه بندی افرازی لازم است که تعداد خوشه دست خوش تغییر و تحولاتی شوند . بعنوان مثال یک خوشه ، وقتی دارای تعداد زیادی نمونه و دارای واریانس غیر معمولی بزرگ همرا با صفات متنوع است به دو بخش تبدیل می شود دو خوشه ای که مراکز شان به اندازه کافی بهم نزدیک اند یکی می شوند داده ها ی پرت و دور افتاده نیز باید شناسایی و با آنها به طرز مناسبی رفتار شود ، مثلا یک داده دور افتاده در شکل زیر باعث شده تا دو خوشه ای که خوب از هم جدا هستند در یک خوشه قرار بگیرند.



شکل(۵) : تاثیر داده دور افتاده در خوشه بندی

همگرایی (گام چهارم)

الگوریتم چه موقع باید متوقف شود ؟ الگوریتم افرازی وقتی تابع معیار نتواند بهبود پیدا کند باید متوقف شود ، یعنی بین دو تکرار متوالی بر چسب خوشه برای همه نمونه ها بدون تغییر بماند.

هیچ ضمانتی برای رسیدن به یک مینیمم کلی برای یک الگوریتم تکراری وجود ندارد ولی باید شروطی را قرار داد تا الگوریتم به تکرارخود ادامه ندهد ، مثلا تعداد تکرارها را از قبل معین کرد یا اگر مربعات خطا از یک حدی کمتر شد حلقه تکرار متوقف شود.

مثال: با استفاده از روش K- میانگین و ماتریس الگوی زیر دو خوشه بدست آورید.

جدول (۱۰) : ماتریس الگو

	صفت		
		اول	دوم
نمونه	x_1	۰	۲
	x_2	۰	۰
	x_3	۱/۵	۰
	x_4	۵	۰
	x_5	۵	۲

ابتدا دو خوشه $C_1 = \{x_1, x_2, x_4\}$ و $C_2 = \{x_3, x_5\}$ را بطور تصادفی انتخاب می کنیم.مرکزیت این دو خوشه عبارتند از:

$$c_1 = \left\{\frac{0+0+5}{3}, \frac{2+0+0}{3}\right\} = \{1.66, 0.66\}$$
$$c_2 = \{\frac{1.5+5}{2}, \frac{0+2}{2}\} = \{3.25, 1.00\}$$

اختلاف درون خوشه ها

$$e_1^2 = [(0-1/66)^2 + (2-0/66)^2] + [(0-1/66)^2 + (0-0/66)^2] + [(5-1/66)^2 + (0-0/66)^2] = 19/36$$
$$e_2^2 = [(1/5-3/25)^2 + (0-1)^2] + [(5-3/25)^2 + (2-1)^2] = 8/12$$

مجموع مربعات کل

$$E^2 = e_1^2 + e_2^2 = 19.36 + 8.12 = 27.48$$

جدول (۱۱) : فاصله نمونه ها از مراکز خوشه ها در تکرار اول

نمونه	فاصله از مرکز خوشه C1	فاصله از مرکز خوشه C2	نزدیک ترین خوشه
x_1	۲/۱۴	۳/۴	c_1
x_2	۱/۷۹	۳/۴	c_1
x_3	۰/۸۳	۲/۰۱	c_1
x_4	۳/۴۱	۲/۰۱	c_2
x_5	۳/۶	۲/۰۱	c_2

حال خوشه های جدید عبارتند از $C_1 = \{x_1, x_2, x_3\}$ و $C_2 = \{x_4, x_5\}$

مرکزیت دو خوشه $\begin{cases} c_1 = (0.5, 0.67) \\ c_2 = (5,1) \end{cases}$

اختلاف درون خوشه ها $\begin{cases} e_1^2 = 4.17 \\ e_2^2 = 2.00 \end{cases} \Rightarrow E^2 = 6.17$

جدول (۱۲) : فاصله نمونه ها از مراکز خوشه ها در تکرار دوم

نمونه	فاصله از مرکز خوشه C1	فاصله از مرکز خوشه C2	نزدیک ترین خوشه
x_1	۱/۴۲۱	۵/۱	c_1
x_2	۰/۸۳۶	۵/۱	c_1
x_3	۱/۲۰۴	۳/۶۴	c_1
x_4	۴/۵۵	۱	c_2
x_5	۴/۶۹	۱	c_2

در این جدول دیده می شود اعضای خوشه های C_1 و C_2 تغییر نمی کنند و معیار همگرایی بدست آمده است.