



به نام خدای بخشنانده مهربان

In the name of Allah, the Beneficent, the Merciful.

کارگاه

مفروضات و مراحل آنالیز خوشه‌ای با تأکید بر کاربرد سنجه Cophenetic

محسن ابراهیم پور
دانشیار موسسه

مرداد ۱۴۰۲

بر پوست کشیده هامون تفتیده
آیه های دلاوریت نوشتی
مزرعه ای شد
با خوشه های سرخ
و جهان یک مزرعه شد
با خوشه خوشه خون

هر که گوید او منم او من نشد خوشه ی او لایق خرمن نشد
من مشو از من بشو تا من شوی خوشه شو تا لایق خرمن شوی

An intelligent being cannot treat every object it sees as a unique entity unlike anything else in the universe. It has to put objects in categories so that it may apply its hard-won knowledge about similar objects encountered in the past, to the object at hand.

یک موجود هوشمند نمی‌تواند با هر شیئی که می‌بیند به عنوان یک موجود منحصر به فرد نامشابه با هر چیز رفتار کند. دیگر در کیهان باید اشیاء را در دسته بندی قرار دهد تا بتواند دانش خود را که به سختی به دست آورده است در مورد اشیاء مشابهی که در گذشته با آنها مواجه شده است، در مورد شیء مورد نظر اعمال کند. استیون پینکر، ذهن چگونه کار می‌کند، 1997.

مقدمه

تشریح و تبیین مفروضات و مراحل کاربرد آنالیز خوشه‌ای و شیوه تحلیل اعتبار و مفهوم یابی خوشه‌ها موضوع این کارگاه است. در ساحت هستی‌شناسانه تنها مفهوم مستقل از مقولات همانا «وجود» است و معرفت بشری و دغدغه بشر بدوی همان قاطب گوریاست و اساس هم تشابهات و تمایزات است. تفاوت متزاید به مفهوم متنازع است.

Factoring ، Clustering، Classification
Component Analysis

و تشخیص، تفاوت، تمایز، تشابه، تباین

مبانی و فرایند شناختی آنالیز خوشه‌ای

خوشه‌یابی فرایند تمایز و تعمیم وجوه مشترک (حداقل دو حیث) مشهود مدرک منتزع تجربه (حواس پنجگانه)، ادراک، تفهم (بر مبنای مقولات اولیه، ثانویه و فاهمه، تعقل و مجرد) و تأیید معیاری است.

البته ادراک مستقل از اندازه‌گیری، ماهیت، هویت و تعریف نیز دستاورد است.

مهمتر از صدها روش آنالیز خوشه‌ای همانا اعتبار و مفهوم‌یابی است. به ویژه در سازه‌ها

طبقه بندی و خوشه یابی میراثی است که از حکمت و فلسفه به معرفت و علم رسیده است و برخی علم را حدس و ابطال پایان ناپذیری صحت و دقت طبقه بندی تشابهات و تفاوت ها تعریف کرده اند و در ساخت پژوهشی تحقیقات نجوم، دیرینه شناسی و نظامی در گسترش فنون آماری و روش های بسیار پیچیده رده بندی ها، طبقه گذاری ها و خوشه یابی ها کارکرد داشته است. بازگشت به نظم های طبیعی، یا شعورهدایت گرایانه و همچنین دقت در امور نظامی به رشد کوانتومی گذار از مقولات پیشینی و کلاسیک نگری کمک کرد و ظهور رایانه های نسل اخیر است که امکان تحلیل های خوشه ای را فراهم ساخته اند.

برسازنده‌های نظری

- خوشه‌های نجومی،
- فضایل سقراطی
- دوالیزم افراطی (قدیم، شرویماری و رنج)
- هفت گانگی قرون وسطایی
- هیجده گانه‌های طبقه‌سازی و مقوله‌بندی
- معما شدگی خوشه‌یابی داروینی
- وفاق سازه‌ای

در پارادایم دانش بر بام است، هر آنچه که لازم است، نیاز است که انسان بداند و بشناسد در متون آمده است و تنها انتقال درست وظیفه است. در پارادایم این همانی، مقولات متشابه، متباین، من وجه و مطلق منطق (هستن) دارند.

در پارادایم این نه آنی، مقولات در هم تنیده و بی نهایت شدن هستند

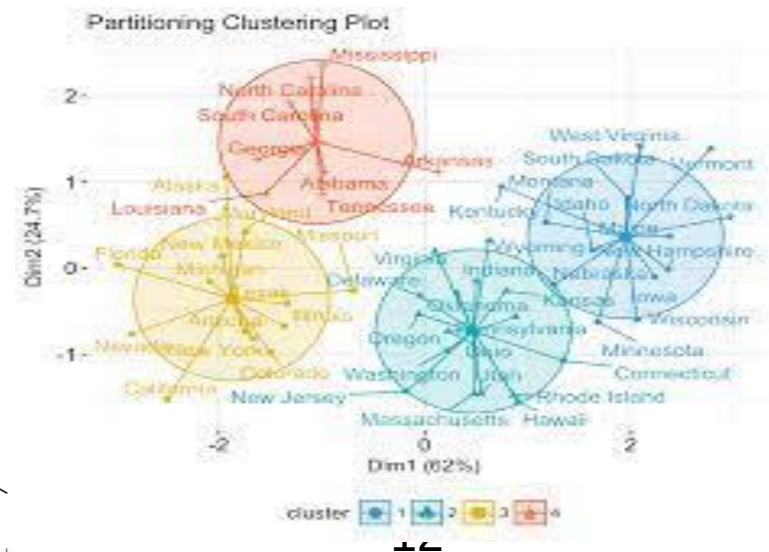
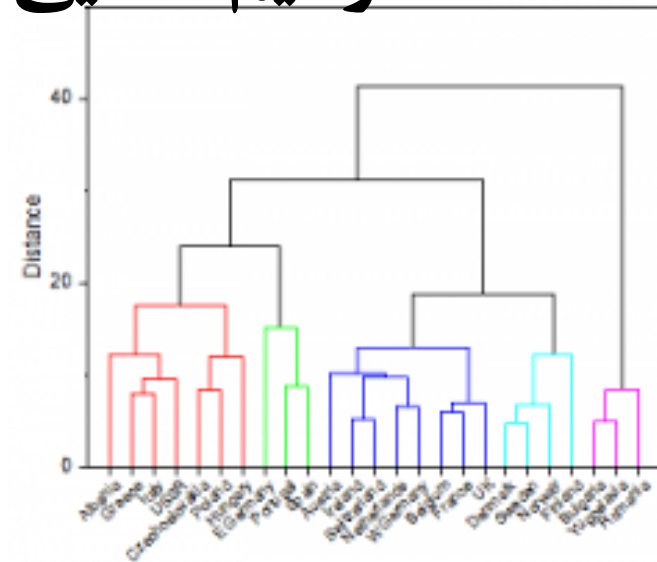
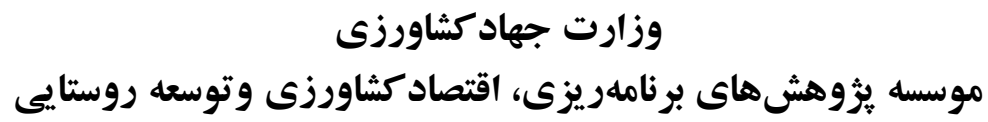
در ساحت Science و شیوه‌های مقوله‌یابی
مقولات فاهمه، مقولات پیشینی، مقولات پسینی، تفاضلات و تأییدات (اولبک تا هابرماس)
الف) پارادایم نیوتنی
مقولات پوزیتیو است
مقولات تفهیم است
مقولات حجیت است
مقولات وفاق پراگماتیستی است

ب) پارادایم کوانتومی
مقولات اکتشافی است
مقولات اینجا و اکنونی است
مقولات بی‌نهایت است
مقولات معقولات است

مقوله یابی در دوره کلاسیک و عصر

در این دوران که حداقل تا عصر حاضر است. مقوله سازی ها (خوشه پروین) حجیتی و مدرسی از خاک به افلاک بود. مقولات دو گانه و سه گانه متباین (ثنویت و تثلیث) جهات اربعه، هفت مقوله، چهل مجموعه بیشترین مفهوم سازی و مصداق یافتگی را دارا هستند. در این گارگاه حسب سنت حسنه مألوف یعنی ذکر نامی و کلامی به عین القضاات همدانی اشارت می رود. در مقوله بندی و مجموعه نگری به مثل (معرفت، مینایی و معاونت) و شهاب الدین (حسن، حب و حزن) در نظریات شهیدی شهیر ماناست (جوهر، عرض و جسم)

ما مرگ و شهادت از خدا خواسته ایم و آن هم به سه چیز کم بها خواسته ایم
کردوست چنین کند که ما خواسته ایم ما آتش و نفت و بویا خواسته ایم.



- تحلیل خوشه‌ای یکی از روش‌های عمده رده‌بندی یا طبقه‌بندی مشاهدات (classifying) نیز هست. اصولاً classify یا رده‌بندی مشاهدات، به منظور تقسیم مشاهدات به گروه‌های متجانس مورد استفاده قرار می‌گیرد به گونه‌ای که ممکن است مشاهدات هر گروه به یکدیگر شبیه باشند در یک طبقه و رده قرار گیرد و یا و گروه‌های مختلفی که کمترین شباهت را با هم داشته باشند در رده‌بندی دیگری قرار گیرند. روش‌های عمده رده‌بندی عبارتند از:
- الف) k-means cluster: این روش هنگامی اعمال می‌شود که تعداد گروه‌ها از قبل معین بوده و برای دستیابی به گروه‌بندی مشاهدات همگن اعمال می‌شود، مانند: گروه‌های عمده شغلی در تجزیه و تحلیل مشاغل روستایی و کشاورزی ایران.
- ب) hierarchical cluster analysis: تحلیل خوشه‌ای سلسله‌مراتبی را زمانی اعمال می‌کنند که تعداد گروه‌ها یا خوشه‌ها از قبل معلوم نباشد.

ج) discriminant analysis: تحلیل تمایزات برای گروه بندی
مشاهدات در یکی از چندین گروه معلوم اعمال می شود.

د) Tree

ه) Twostep cluster

ز) Nearest Neighbor

- تفاوت روشهای تحلیل خوشه ای با تحلیل ممیزی (تمایزی) را می توان در موارد زیر بیان نمود:
- الف) برای رده بندی (گروه بندی) مشاهدات جدید استفاده از تحلیل خوشه ای امکان پذیر است

زیرا تحلیل تمایزات (توابعی بر حسب مقادیر متغیرهای مورد اندازه‌گیری)، امتیازی را برای عضویت مشاهدات جدید در هر گروه محاسبه میکند و با ورود هر مشاهده جدید می‌توان با محاسبه امتیاز آن بر حسب هر یک از توابع، وضعیت عضویت آن را در یکی از گروه‌ها تعیین کرد.

ب) در تحلیل تمایزات باید از هر یک از گروه‌های مورد نظر، نمونه‌های مناسبی در اختیار داشت. اما در تحلیل خوشه‌ای اعضاء مشخصی از مشاهدات به گروه‌های معین منتسب نموده و بنابراین ضرورت وجود نمونه‌های مناسب در هر گروه منتفی است.

بنابراین اعمال روش‌های رده‌بندی (classifying)، بر حسب معلوم بودن یا نبودن تعداد گروه‌ها، تعداد گروه‌های هدف، نوع متغیرهای مورد اندازه‌گیری متفاوت خواهد بود و روش‌های خوشه‌بندی یا تحلیل خوشه‌ای نیز بر حسب معلوم بودن یا نبودن تعداد گروه‌های مشابه، اهداف خوشه‌بندی، نوع خوشه‌بندی متنوع و متفاوت هستند. همچنین شیوه خوشه‌بندی بر حسب روش‌های مختلفی که عمده‌ترین آنها تشریح خواهند شد، متفاوت می‌باشد.

مراحل اساسی آنالیز خوشه

شش مرحله اساسی تحلیل خوشه‌ای، یکی پس از دیگری و بصورت منظم و بترتیب بیان می‌گردند:

اولین مرحله و سرآغاز تحلیل خوشه‌ای، بدست آوردن ماتریس داده‌ها می‌باشد.

ستون‌های ماتریس داده‌ها به موضوعات (objects) و سطرهاى آن به مقادیر نسبت‌های هر یک از موضوعات اختصاص می‌یابد. در واقع ستونهای ماتریس داده‌ها در تحلیل خوشه‌ای به موضوعاتی اختصاص می‌یابد که می‌خواهیم شباهت‌های (عدم شباهت‌ها) آنها را با یکدیگر برآورده نمائیم و سطرهاى آن در بردارنده نسبت‌ها یا مقادیر (attributes) هر یک از موضوعات است.

مثال داده

انتخاب گروه بر حسب دو مشخصه (نسبت)

گروه‌های شغلی ویژگی ها	گروه شغلی خدمات	گروه شغلی صنایع روستایی	گروه شغلی شیلات	گروه شغلی دامداری	گروه شغلی گرفغل
(نیازهای آموزش)	5	30	30	20	10
(شرایط احراز شغل)	10	15	10	20	5

0	3	4	2	1	5
3	-	-	-	-	-
4	5	-	-	-	-
2	14/1	11/2	-	-	-
1	20/6	22/4	18	-	-
5	25	25/5	18	7/7	-

Rearranged resemblance matrix

0	3	4	2	1	5
3	-	-	-	-	-
4	5	-	-	-	-
2	14/1	11/2	-	-	-
1	20/6	22/4	18	-	-
5	25	25/5	18	7/7	-

ماتریس باز تنظیم شده
بر اساس خوشه ۳ و ۴ - ۱ و ۵ - ۲ و ۱ - ۵ و ۲ با ۳ و ۴

30	30	20	10	5
10	15	20	5	10

Cophenetic Matrix

0	1	2	3	4	5
1	-	-	-	-	-
2	21/6	-	-	-	-
3	21/6	12/7	-	-	-
4	21/6	12/7	5	-	-
5	7/07	21/6	21/6	21/6	-

ر جهت محاسبه cophenetic مقادیر ماتریس تشابه و ماتریس

زوج ها	X	Y	X Y	X ₂	Y ₂
(1و2)	18	21/6	18×21/6	(18) ²	(21/6) ²
(1و3)	20/6	21/6	20/6×21/6	(20/6) ²	(21/6) ²
(1و4)	22/4	21/6	22/4×21/6	(22/4) ²	(21/6) ²
(1و5)	7/07	7/07	7/07×7/07	(7/07) ²	(7/06) ²
(2و3)	14/1	12/7	14/1×12/7	(14/1) ²	(12/7) ²
(2و4)	11/2	12/7	11/2×12/7	(11/2) ²	(12/7) ²
(2و5)	18	21/6	18×21/6	(18) ²	(12/6) ²
(3و4)	5	5	5×5	5 ₂	5 ₂
(3و5)	25	21/6	25×21/6	25 ₂	(21/6) ²
(4و5)	25/5	21/6	25/5×21/6	(25/5) ²	(21/6) ²
	Σ	166/9	167/1	3248/6	3196/9

$$R_{XY} = \frac{R_{PXY}}{\sqrt{SSX \cdot SSY}} = \frac{3193/5 - \frac{(166/9 \times 167/1)}{10}}{\sqrt{\left[3248/6 - \frac{866/912}{10}\right] \left[3196/9 - \frac{(167/1)^2}{10}\right]}} = 0/92$$

روش‌های خوشه‌بندی

علاوه بر "روش غیروزی زوجی-گروهی با بکارگیری میانگین‌های حسابی" که در مثال مورد استفاده قرار گرفت و با اعمال آن به خوشه‌بندی پرداخته شد، روشهای دیگری نیز وجود دارد که در اینجا به تشریح عمده‌ترین و معمول‌ترین آنها می‌پردازیم البته فوق (UPGMA)، در تحلیل خوشه‌ای سلسه مراتبی، بیشترین کاربرد را داشته و معمولاً تحلیل گران خوشه‌ای از آن استفاده می‌کنند. روش‌های دیگری که جنبه عام‌تر و معمول‌تری دارند عبارتند از:

(between groups linkage): روش ارتباط بین گروهی

در این روش، ترکیب خوشه‌ها از کمینه کردن متوسط فاصله بین تمام زوج‌های مشاهداتی که در خوشه‌های مختلف قرار دارند، ایجاد می‌شوند. در این روش از کلیه فواصل موجود بین نقاط خوش‌ها استفاده می‌شوند نه فقط نزدیکترین یا دورترین فاصله‌ها.

این روش همان روش (UPGMA) می‌باشد که در مشاغل عددی اعمال گردید.

(within groups linkage): روش ارتباط درون گروهی-3

در این روش خوشه‌بندی بگونه‌ای صورت می‌گیرد که معدل فاصله نقاط درون خوشه‌ها از یکدیگر به حداقل (کمترین مقدار) برسد.

روش ارتباط یگانه یا نزدیکترین همیاسه‌ها {nearest neighbor (simple linkage)}:

در این روش، کمترین فاصله هر مشاهده رانسبت به مشاهدات موجود بدست آورده و سپس این عمل را برای کلیه مشاهدات باقی‌انده انجام داده و نسبت به کلیه عناصر خوشه‌های موجود تکرار می‌کنیم.

– روش دورترین همسایه ها (furthest neighbor):

در این روش فاصله بین دو خوشه بر حسب فاصله بین دورترین نقاط آن محاسبه شده و بر مبنای مقادیر حاصل خوشه‌بندی صورت می‌گیرد.

(centroid clustering): روش خوشه‌بندی متمرکز-3

در این روش فاصله بین دو خوشه، فاصله بین میانگین‌های آن می‌باشد و فاصله‌ای که خوشه‌ها با هم ترکیب می‌شوند از مرحله‌ای به مرحله دیگر کاهش می‌یابند.

(median clustering): روش خوشه‌بندی میانه-3

در این روش، به دو خوشه‌ای که ترکیب می‌شوند وزن‌های یکسانی صرف نظر از تعداد نقاط آنها داده می‌شود. این عمل سبب می‌شود که گروه‌های کوچک نسبت به سایر گروه‌های دیگر اثر مشابهی در ساختن خوشه‌های بزرگتر داشته باشند.

روش وارد (ward's):

در این روش ابتدا میانگین متغیرها در داخل هر خوشه محاسبه شده و سپس برای هر مشاهده مربع فاصله اقلیدسی میانگین‌های خوشه‌ها محاسبه می‌گردد. این فاصله برای تمامی مشاهدات جمع شده و در مرحله خوشه‌بندی، در هر مرحله، دو خوشه به گونه‌ای ترکیب می‌شوند که کوچکترین افزایش در مجموع مربعات فواصل داخل خوشه‌ای را داشته باشند.

روش‌های فاصله‌ای (interval):

در این روش‌ها معیار تشابه یا عدم تشابه بین مشاهدات مختلف بر حسب میزان فاصله بین دو نقطه اندازه‌گیری می‌شود که بهترین این معیارها عبارتند از:

الف) معیار فاصله اقلیدسی: در مثال عددی ما، از این معیار جهت دستیابی به فاصله‌های بین نقاط استفاده شد و فرمول آن عبارت بود از:

$$d_{jk} = D = \sqrt{\sum_{i=1}^n (x_{ij} - x_{ik})^2} = \left[\sum_{i=1}^n (x_{ij} - x_{ik})^2 \right]^{1/2}$$

ب) معیار فاصله اقلیدسی مربع: معیار مربع فاصله اقلیدسی همانند معیار بالاست جز اینکه از جذر مقادیر صرف نظر می‌شود، یعنی:

$$D^2 = \sum_{i=1}^n (x_{ij} - x_{ik})^2$$

(ب) معیار کسینوس (Cos): در صورتی که دو بردار متعامد باشند، کسینوس زاویه معادل صفر است. مقدار این معیار بین $+1$ و -1 است.

(ج) معیار پیرسون (pearson): ضریب همبستگی خطی بین مقادیر دو بردار.

(د) معیار chebychev: در این معیار، فاصله بین دو مشاهده برابر ماکزیمم قدر مطلق اختلافات مقادیر در هر متغیر تلقی می‌شود.

$$cheb = \max |x_{ki} - x_{kj}|$$

(ه) معیار بلوک (Block): معیار فاصله در این روش، معادل مجموع قدر مطلق انحرافات مقادیر برای هر متغیر است.

$$Block = \sum |x_{ij} - x_{ik}|$$

،یگر روش‌ها فاصله‌ای که کمتر معمول می‌باشند، عبارتند از: customized و minkowski

(ب) روش‌های شمارشی (counts):

مهمترین روش‌های محاسباتی آن عبارتند از:

الف) معیار مربع کای اسکوار (χ^2) یا معیار عدم تشابه بر اساس مربع χ^2 بزرگی این مقدار به مجموع فراوانی

دو مشاهده یا متغیری که عدم تشابه محاسبه می‌شود بستگی دارد. مقادیر مورد انتظار از مدلی که

استقلال مشاهدات یا متغیرها را فرض می‌کند، به دست می‌آید. مقدار اندازه ریشه دوم آمار χ^2 است.

ب) معیار χ^2 : این اندازه (معیار) سعی می‌کند اندازه نمونه را برای کاهش اثر مقادیر فراوانی‌های مشاهده

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

شود واقعی مقادیر محاسبه کند. χ^2 مقدار از χ^2 را با کل فراوانی نرمالیزه می‌کند

روش‌های باینری خوشه‌یابی

روش‌های دودویی (Binary):

برای استفاده از این روش‌ها ابتدا مقادیر را با اعمال تبدیل‌های خاص به صورت دودویی (Binary) در می‌آوریم و آنگاه معیارهای فاصله‌ای مورد نظر را اعمال می‌کنیم؛ این معیار عبارتند از:

الف - معیار فاصله اقلیدسی باینری.

ب - مربع فاصله اقلیدسی باینری

ج - معیار اندازه تفاوت حجم.

د - معیار اندازه تفاوت الگوها

معیار اندازه‌های عدم تشابه واریانس و شکل.

ضریب غیر متریک بری - کورسیت

ادامه خوشه بندی

با این وجود بنظر می‌رسد که ضریب فاصله اقلیدسی با مبانی مثلثاتی و ریاضی قابل فهم و ساده و با ارائه اندازه‌های واقعی فاصله‌ها بین نقاط، بهترین معیار فاصله باشد و روش خوشه بندی ارتباط بین گروهی یا روش غیروزی گروهی - زوجی با اعمال میانگین‌های حسابی"، بهترین روش خوشه بندی و قابل فهم ترین آنها بوده‌است، ولی برای سطوح اندازه گیری فاصله ای و بالاتر کاربرد دارند.

آنالیز خوشه ای و عاملی همراه با رگرسیون یک نظام جامع آماری هستند.

پارادایم‌های نوین خوشه‌یابی و طبقه‌بندی

(الف) نومی‌نالیسم (هو)

(ب) امپریسم (آمار و تجربه)

(ج) مثبت‌نگری

(د) تاریخت (تا هابرماس)

(ه) تامیت‌انگاری

(و) بازاندیشانه



با سپاس از توجه شما