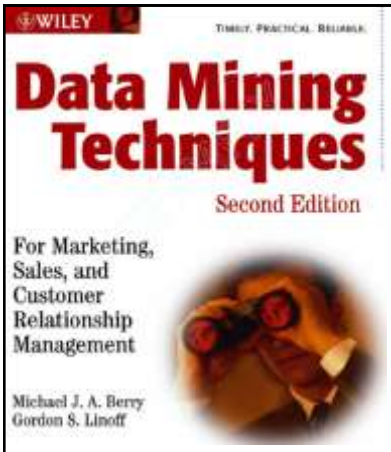


داده کاوی

تحلیل خوشه ای

سمیه علیزاده

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مطالب مورد بحث در کلاس

- تعاریف و مفاهیم
- انباره داده ها
- آماده سازی داده ها (پیش پردازش داده ها)

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مطالب مورد بحث در کلاس

- خوشه بندی
- دسته بندی
- قوانین انجمنی
- سریهای زمانی
- وب کاوی
- متن کاوی
- پیوندکاوی و تحلیل شبکه های اجتماعی

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مطالب مورد بحث در کلاس

- متدولوژی اجرای پروژه های داده کاوی
- کاربردهای داده کاوی در بازاریابی
- کاربردهای داده کاوی در مدیریت ارتباط با مشتری
- امنیت در داده کاوی

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

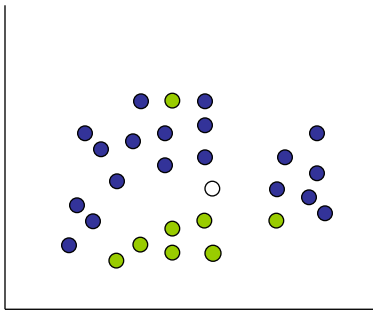
خوشه بندی

- تجزيه و تحليل خوشه‌اي روشي براي گروه بندي داده‌ها يا مشاهدات با توجه به شباهت يا درجه نزديكي آنها است.

- از طريق تجزيه و تحليل خوشه‌اي ، داده‌ها يا مشاهدات به دسته‌هاي همگن و متمايز از هم تقسيم مي‌شوند.

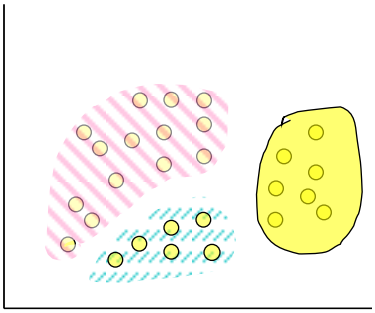
خوشه بندی

- درروش خوشه بندی هيچ دسته اي از قبل وجود ندارد و در واقع متغيرها بصورت مستقل و وابسته تقسيم نمي شوند. بلكه ما در اينجا بدنبال گروه هايي از داده ها هستيم كه به هم شباهت دارند و با كشف اين شباهت ها مي توان رفتارها را بهتر شناسايي كرد و بر مبناي آنها طوري عمل كرد كه نتيجه بهتري حاصل شود.



خوشه بندی

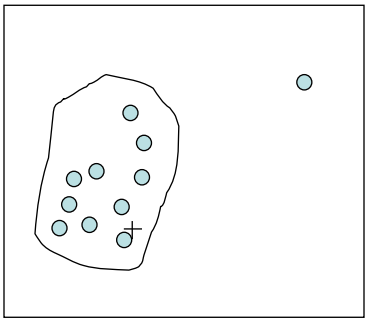
- درروش خوشه بندی هيچ دسته اي از قبل وجود ندارد و در واقع متغيرها بصورت مستقل و وابسته تقسيم نمي شوند. بلكه ما در اينجا بدنبال گروه هايي از داده ها هستيم كه به هم شباهت دارند و با كشف اين شباهت ها مي توان رفتارها را بهتر شناسايي كرد و بر مبناي آنها طوري عمل كرد كه نتيجه بهتري حاصل شود.



سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

كاربردهای خوشه بندی در آماده سازی داده ها

- در بعضي موارد از خوشه بندي براي داده هايي كه با ساير داده ها تفاوت چشمگير دارد استفاده مي‌نمايند.
- مثلا يكسري از مشتريان همگي خريدي بالاي 100 دلار در ماه دارند به غير از يكي



سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

نقاط قوت روش خوشه بندی

- **روش خوشه بندي يك روش غير مستقيم است.**
  - بدين معني كه اين روش را مي توان حتي هنگامي كه هيچ نوع اطلاعات قبلي از ساختار داخلي پايگاه داده ها نداريم استفاده نمود. از اين روش مي توان براي كشف الگوهاي پنهان و بهبود عملكرد روشهاي مستقيم نيز استفاده نمود.
- **خوشه بندي را مي توان براي داده هاي گوناگون استفاده نمود.**
  - با انتخاب درست اندازه فاصله هاي گوناگون خوشه بندي را مي توان براي بيشتر انواع داده ها استفاده نمود.
- **استفاده از اين روش آسان است.**

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

نقاط ضعف روش خوشه بندی

- انتخاب اندازه هاي دقيق فواصل و وزنها كار آساني نمي باشد.
- اين روش به پارامترهاي اوليه نظير تعداد خوشه ها ، حداقل نزديكي و خوشه هاي اوليه حساس است.
- تفسير نتايج اين روش مي تواند مشكل باشد و معمولا نياز به تحليل افراد خبره در زمينه تجارت مورد نظر دارد.

سميه عليزاده هيات علمی دانشكده صنايع
دانشگاه خواجه نصير طوسی

یک خوشه بندی خوب چیست؟

- يك روش خوشه بندي خوب خوشه‌هاي با كيفيت بالا براساس دو معيار زير را توليد مي‌كند:

شباهت بالاي نقاط داخلي‌هر كلاس و شباهت كم بين نقاط كلاسهاي‌مختلف .

- كيفيت نتايج خوشه بندي بستگي به روش اندازه‌گير شباهت به كار رفته و همچنين پياده سازي آن روش دارد

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

مراحل تجزیه و تحلیل خوشه‌ای

- انتخاب معيار شباهت يا نزديكي مشاهدات
- انتخاب روش تجزيه و تحليل خوشه‌اي
- تصميم‌گيري در مورد تعداد خوشه‌ها
- تفسير دسته‌ها يا گروه‌هاي تشكيل شده

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

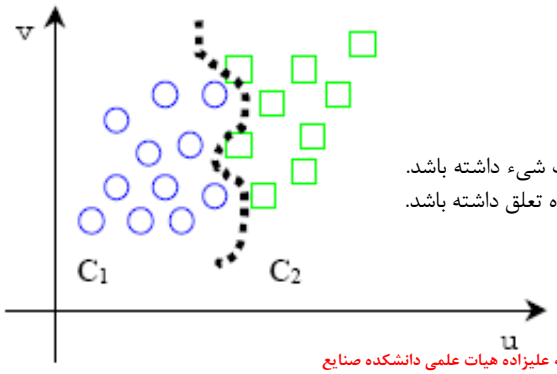
رویکردهای اصلی خوشه بندی

- روش افراز بندي
- روش سلسله مراتبي
- روش مبتني برچگالي
- روش Grid-based
- روش مبتني بر مدل

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

روش افزار بندی

فرض كنيد يك پايگاه داده از n شيء داشته باشيم. يك روش افرازبندي ، K افزار از اين داده‌ها درست مي‌كند بطوريكه هر افراز يك خوشه را نشان مي‌دهد. پس داده‌ها در k گروه گروه بندي مي‌شوند كه بايد داراي دو شرط زير باشند:



الف) هر گروه بايستي حداقل يک شيء داشته باشد.
ب) هر شي بايد تنها به يک گروه تعلق داشته باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

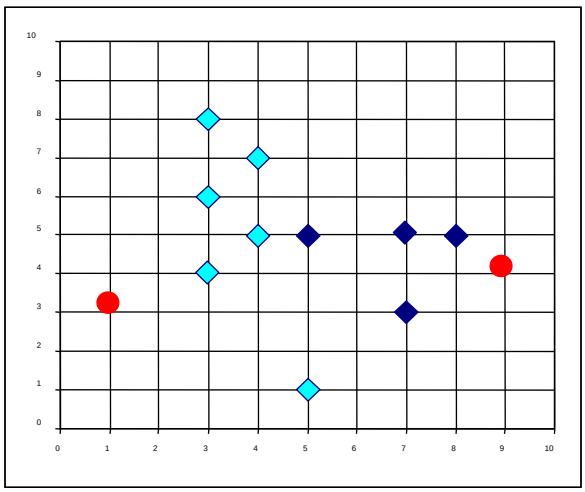
روش افزار بندی

- الگوریتم *K-means* که هر خوشه با میانگین اشیاء آن خوشه ، نمایش داده مي‌شود. (با مرکز خوشه)

- الگوریتم *K-medoids* که هر خوشه با يكي از اشياء كه در نزديكي مركز خوشه جاي گرفته است ، نمايش داده مي‌شود.

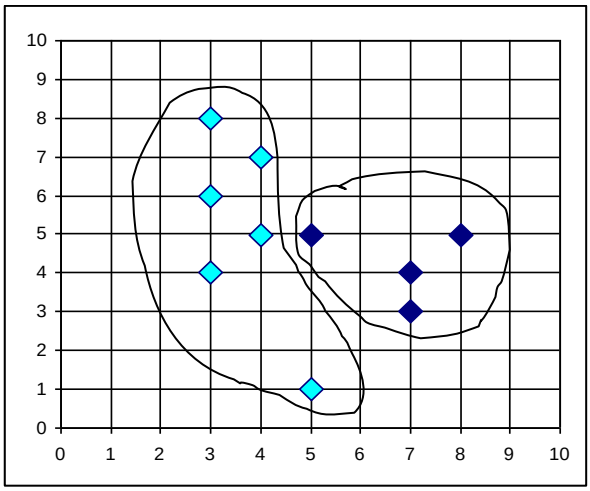
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

انتخاب K عضو دلخواه اولیه



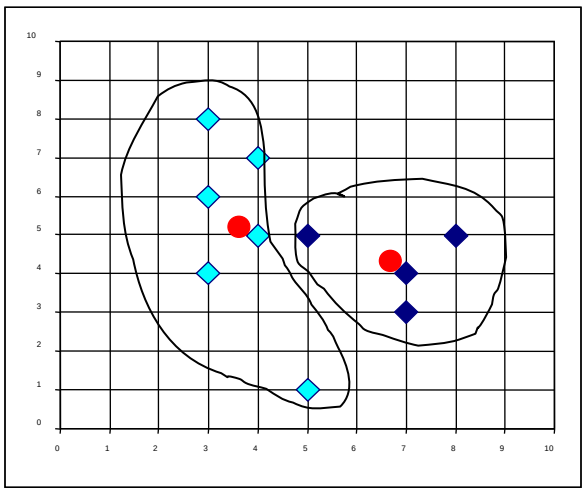
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

هر عضو به شبیه ترین می پیوندد



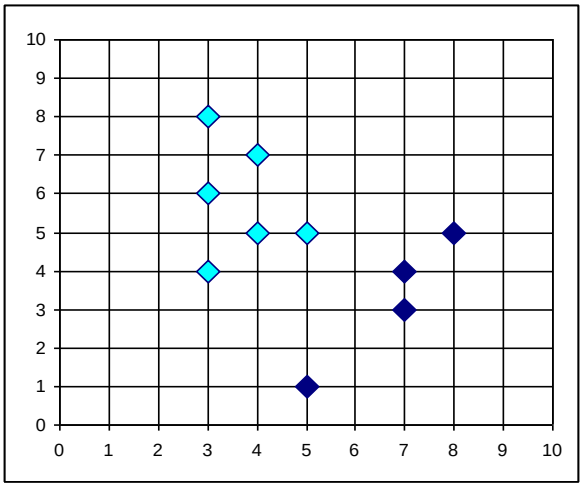
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

به روز کردن میانگین داده ها



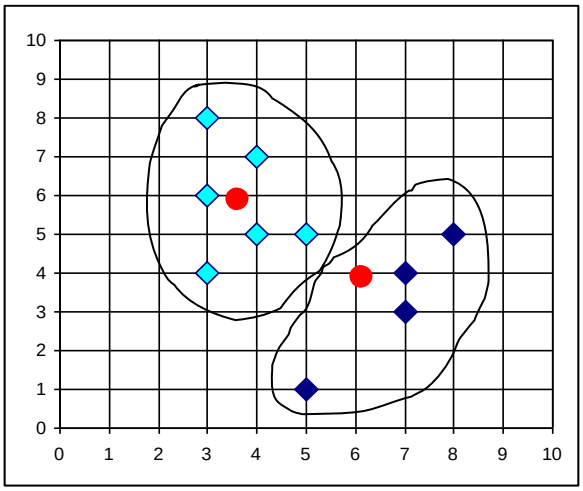
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

بازنگری مجدد



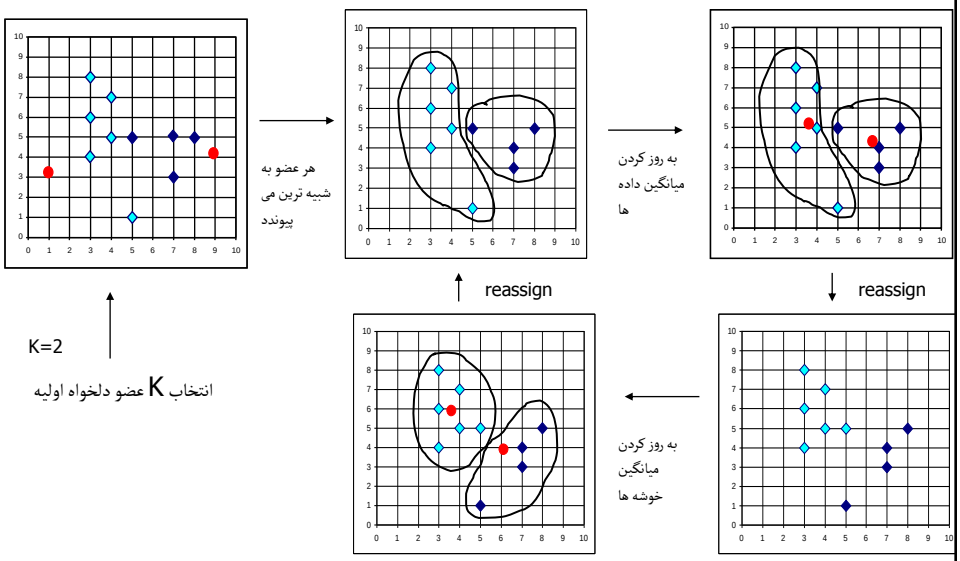
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

به روز کردن میانگین خوشه ها



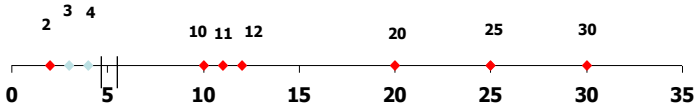
سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

الگوریتم K-means

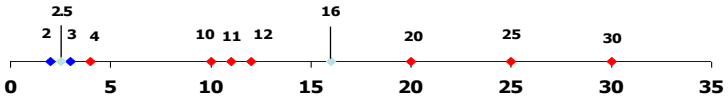


مثال K-Means

- {2,4,10,12,3,20,30,11,25}, k=2
- m_1=3,m_2=4

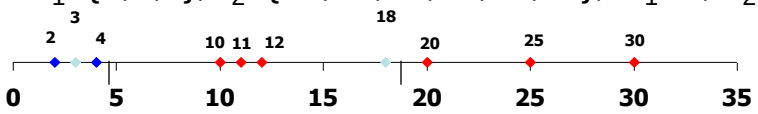


- K_1={2,3}, K_2={4,10,12,20,30,11,25}, m_1=2.5,m_2=16

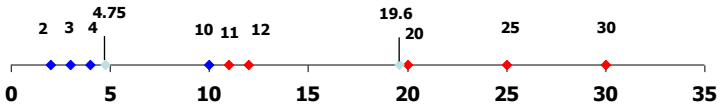


سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- $K_1=\{2,3,4\}, K_2=\{10,12,20,30,11,25\}$, $m_1=3, m_2=18$

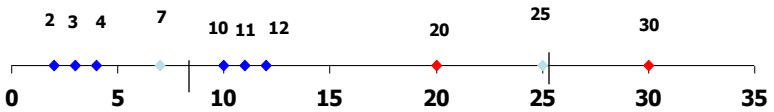


- $K_1=\{2,3,4,10\}, K_2=\{12,20,30,11,25\}$, $m_1=4.75, m_2=19.6$



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

- $K_1=\{2,3,4,10,11,12\}, K_2=\{20,30,25\}$, $m_1=7, m_2=25$



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

K-Means

نقاط قوت

- این روش نسبتاً براي پايگاههاي دادة بزرگ كارا و ارتقا پذير مي باشد زيرا پيچيدگي محاسباتي‌اش عبارتست از O(tkn) كه: n تعداد كل اشياء ، K تعداد خوشه‌ها و t تعداد تكرارهاي الگوريتم است . اين روش اغلب به يك بهينة محلي ختم مي‌شود نه يك بهينة سراسري . بهينه سراسری از طریق روشهایی مانند ژنتیک بدست می آید.

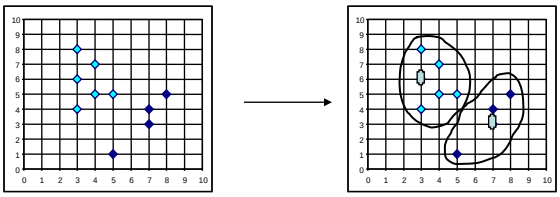
K-Means

نقاط ضعف

- روشK-means تنها هنگامي كاربرد دارد كه بتوان مراكز خوشه‌ها را تعريف نمود. مثلاً براي داده‌هايي با صفات رده‌اي اين روش كارا نيست.
- از معايب اين روش تعيين K می باشد كه مي‌بايست كاربر ابتدا آنرا معين كند و راه خاصي براي تعيين آن مشخص نشده است.
- همچنين اين روش براي كشف خوشه‌هايي با شكلهای پيچيده مناسب نيست.
- يكي از مهمترين نقاط ضعف اين روش اين است كه در برابر نويزها و داده‌هاي دوراز مركز حساس است زيرا اين داده‌ها به راحتي مراكز را تغيير مي‌دهند و ممكن است نتايج مطلوبي حاصل نشود.

روش *K-Medoids*

- K-Medoids: در اين روش بجاي استفاده از مركز يك خوشه به عنوان مرجع، مي‌توان از medoid ها استفاده كرد. يعني شي كه در مركزترين محل يك خوشه مي‌باشد. بنابراين روش افراز هنوز مي‌تواند مبتني بر اصل مينمم سازي مجموع عدم شباهتها ميان هر شي و شيء مرجع متناظرش شكل بگيرد.



PAM (Partitioning Around Medoids, 1987)

استراتژي اساسي آلگوريتم خوشه بندي K-medoids پيدا كردن K شيء نمايندة آغازين (medoid) بطور دلخواه از n شيء پايگاه داده‌ها مي‌باشد.

- هر شيء باقيمانده با medoid اي هم خوشه مي‌شود كه بيشترين شباهت را به آن داشته باشد. سپس اين استراتژي مكرراً يكي از اشياء medoid را با يكي از اشياء غير medoid جايگزين مي‌كند به طوري كه كيفيت نتيجة خوشه‌بندي بهبود بخشيده شود.
- اين كيفيت با بكارگيري تابع هزينه تخمين زده مي‌شود كه ميانگين عدم تشابه بين يك شيء و medoid آن خوشه را اندازه‌گيري مي‌كند.

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

الگوريتم k-medoids

- شيء به عنوان medoid هاي اوليه به صورت دلخواه اختيار كن.
- تكرار كن تا اينكه هيچ تغييري رخ ندهد.
- هر كدام از اشياء باقيمانده را به خوشه‌اي با نزديكترين medoid تخصيص بده
- بطور تصادفي يك شي غير medoid را انتخاب كن ،.
- هزينة نهايي S را از عوض كردن (medoid آن خوشه) و محاسبه كن
- اگر S<0 آنگاه جاى عناصر را عوض كن تا مجموعه K تا medoid جديد شكل بگيرد.

سميه عليزاده هيات علمى دانشكده صنايع
دانشگاه خواجه نصير طوسى

هزينه تغييرات

- براي اندازه‌گيري اينكه شيء بهتر از O به عنوان يك medoid هست يا خير ، كافيست حاصل معادله زير را بدست آوريم و اگر آنگاه O' جابه جايي $E(O') - E(O) < 0$ (ميباشد.
- :

$$E = \sum_{i=1}^{n} \sum_{p \in C_i} d(p, o_i)^2$$

- در اينجا E در اصل ميزان كل فاصله ها از هر نقطه را نشان مي دهد و S ميزان هزينه تعويض مي كرد كه منفي بودن آن برابر سود است.

سميه عليزاده هيات علمى دانشكده صنايع
دانشگاه خواجه نصير طوسى

K-Medoids مثال

- 1, 2, 6, 7, 8, 10, 15, 17, 20 – break into 3 clusters
  - Cluster = 6 – 1, 2
  - Cluster = 7
  - Cluster = 8 – 10, 15, 17, 20
- Random non-medoid – 15 replace 7 (total cost=-13)
  - Cluster = 6 – 1 (cost 0), 2 (cost 0), 7(1-0=1)
  - Cluster = 8 – 10 (cost 0)
  - New Cluster = 15 – 17 (cost 2-9=-7), 20 (cost 5-12=-7)
- Replace medoid 7 with new medoid (15) and reassign
  - Cluster = 6 – 1, 2, 7
  - Cluster = 8 – 10
  - Cluster = 15 – 17, 20

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

K-Medoids مثال

- Random non-medoid – 1 replaces 6 (total cost= -1)
  - Cluster = 8 – 6 (cost 2-0=2), 7 (cost1-1=0), 10 (cost 0)
  - Cluster = 15 – 17(cost 0), 20(cost 0)
  - New Cluster = 1 – 2 (cost 1-4= -3)
- Replace medoid 6 with new medoid (1) and reassign
  - Cluster = 1 – 2
  - Cluster = 8 – 6, 7, 10
  - Cluster = 15 – 17, 20
- Random non-medoid – 10 replaces 8 (total cost=2) don't replace
  - Cluster = 1– 2(cost 0)
  - Cluster = 15 – 17 (cost 0), 20(cost 0)
  - New Cluster = 10 – 6 (cost 0), 7 (cost 0), 8 (cost 2-0=2)

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

K-Medoids مثال

- **Random non-medoid – 17 replaces 15 (total cost=0) don't replace**
  - **Cluster = 1 – 2(cost 0)**
  - **Cluster = 8 – 6 (cost 0), 7 (cost 0), 10 (cost 0)**
  - **New Cluster = 17 – 15 (cost 2-0=2), 20(cost 3-5=-2)**
- **Random non-medoid – 20 replaces 15 (total cost=6) don't replace**
  - **Cluster = 1 – 2(cost 0)**
  - **Cluster = 8 – 6 (cost 0), 7 (cost 0), 10 (cost 0)**
  - **New Cluster = 20 – 15 (cost 5-0=5), 17(cost 3-2=1)**
- **Other possible changes all have high costs**
  - **1 replaces 15, 2 replaces 15, 1 replaces 8, …**
- **No changes, final clusters**
  - **Cluster = 1 – 2**
  - **Cluster = 8 – 6, 7, 10**
  - **Cluster = 15 – 17, 20**

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

PAM (Partitioning Around Medoids)

- روش ذكر شده PAM (partitioning around medoids) نام دارد كه يكي از اولين الگوريتمهاي K-medoids است و تلاش مي‌كند k افراز براي n شيء معين كند.
- در اين روش همة زوجهاي ممكن از اشياء آناليز مي‌شوند كه يكي medoid و ديگري غير medoid است.
- يك شيء با شييء جابه جا مي‌شود كه بيشترين كاهش را در خطای مربع داشته باشد.
- لذا اين روش براي پايگاه داده‌هاي بزرگ مشكل است. براي رفع اين اشكال از الگوريتم‌هاي CLARA , CLARANS استفاده مي‌شود.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

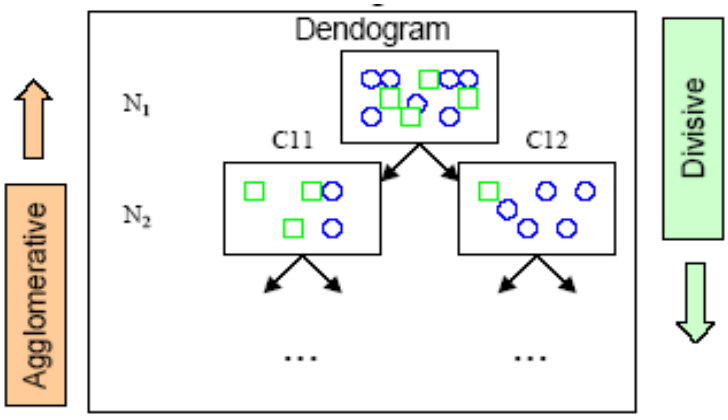
CLARA (Clustering Large Applications) (1990)

- توسط Kaufmann , Rousseeuw در 1990 ارائه شد.

این الگوريتم براي پايگاه داده‌هاي بزرگ بكار مي‌رود. به اين صورت كه چندين نمونة تصادفي از اين پايگاه داده بر مي‌دارد و الگوريتم PAM را روي هر نمونه اجرا مي كند و آن نمونه را خوشه بندي مي‌كند. سپس عناصر باقيماندة پايگاه داده را به نزديكترين خوشه تخصيص مي‌دهد.

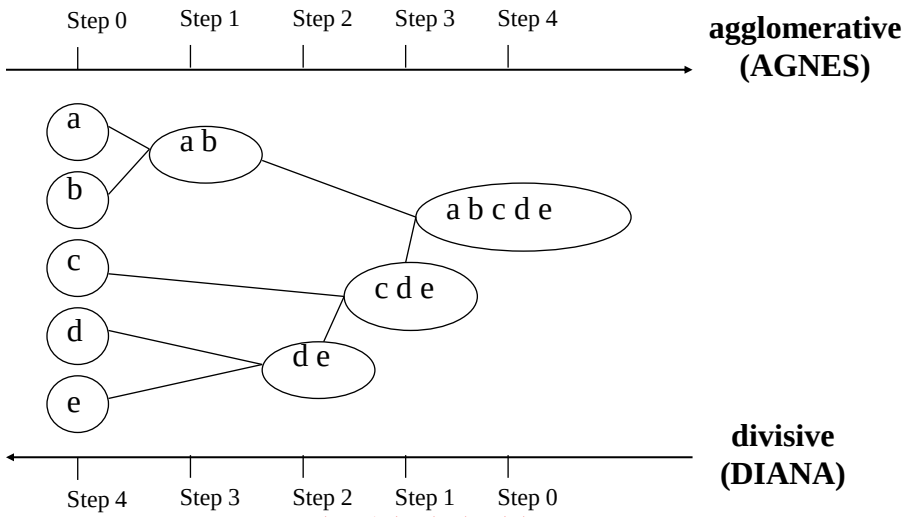
سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

خوشه بندی سلسله مراتبی



سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

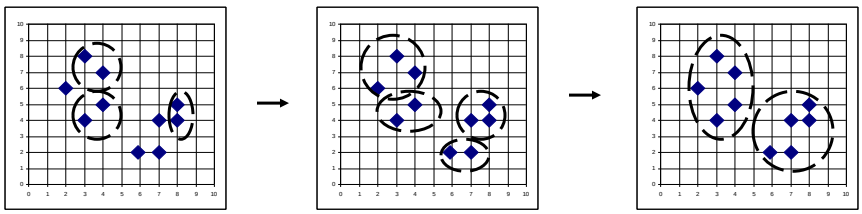
خوشه بندی سلسله مراتبی



سمیه علیزاده هیات علمی دانشکده صنایع

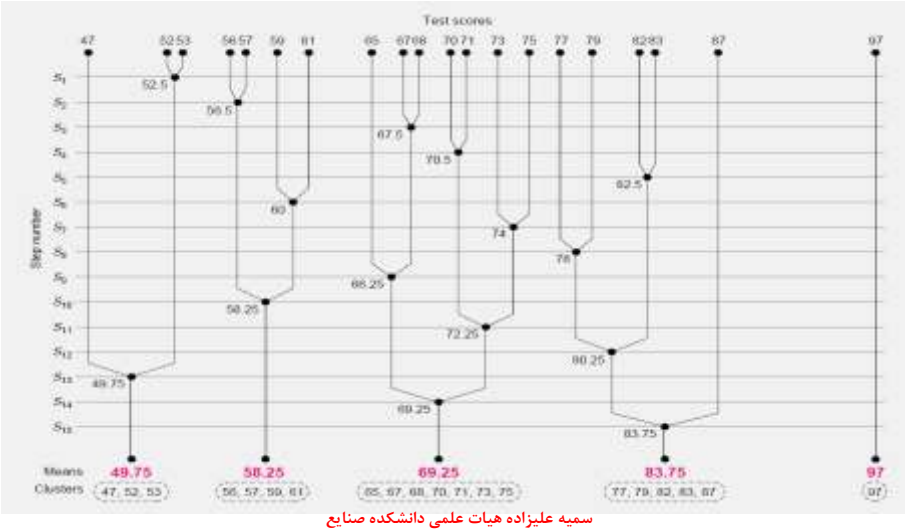
AGNES (Agglomerative Nesting)

- خوشه‌ها را مکرراً با هم ترکیب مي‌کند. به این صورت که ابتدا هر يك از اشیاء را در داخل يك خوشه قرار مي‌دهد و سپس این خوشه‌ها را با ترکیب کردن به خوشه‌هاي بزرگ و بزرگتر تبدیل مي‌کند تا اینکه همة اشیاء در يك خوشه قرار گیرند و یا به شرط پایان برسد.



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Agglomerative Clustering مثال



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

(top-down) Divisive

- خوشه‌ها را مكرراً تقسيم مي‌كند. دقيقاً بر عكس روش تركيبي عمل مي‌كند به اين صورت كه ابتدا همة اشياء در يك خوشه قرار دارندو الگوريتم اين خوشه را به خوشه‌هاي كوچك و كوچكتر تجزيه مي كند تا اينكه هرشيء در يك خوشه قرار گيرد. اين روش معمولاً مناسب نيست و خيلي كم مورد استفاده قرار مي‌گيرد زيرا پيچيدگي محاسباتش بالا ست

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

فاصله در خوشه بندی سلسله مراتبی

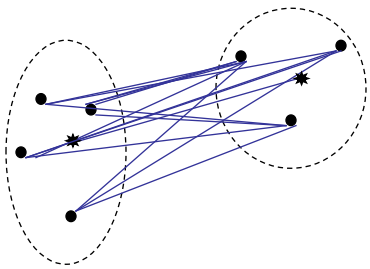
- محاسبه فاصله ها مهم می باشد
- روش محاسبه فاصله نیز مهم است

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

فاصله در خوشه بندی سلسله مراتبی

- معیارهاي گوناگوني که در روشهاي سلسله مراتبي براي فاصلة بين خوشه‌ها بکار مي‌رود ، عبارتند از :

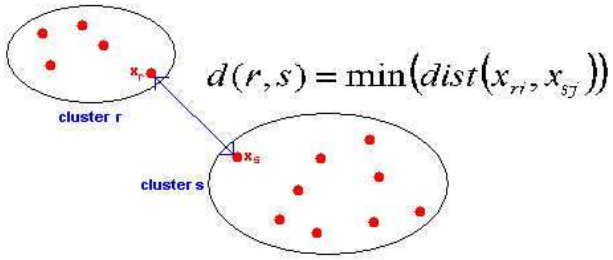
- ***Single Link***
- ***Complete Link***
- ***Average Link***
- ***Centroid***



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

Single linkage

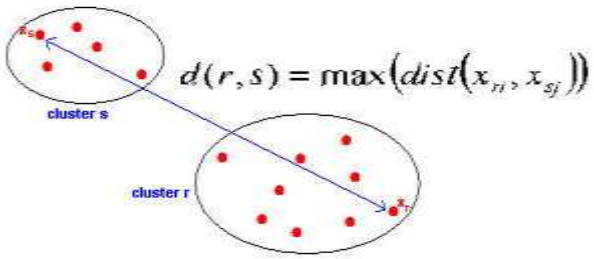
- فاصله بين دسته‌ها بر حسب حداقل فاصله ممكنه بين عناصر آنها محاسبه مي‌شود.
- كليه فاصله بين زوجهاي عناصر دو دسته محاسبه شده و حداقل آنها فاصله بين دو دسته را تعيين مي‌كند.



$$d(r,s) = \min\left(dist\left(x_{ri}, x_{sj}\right)\right)$$

Complete linkage

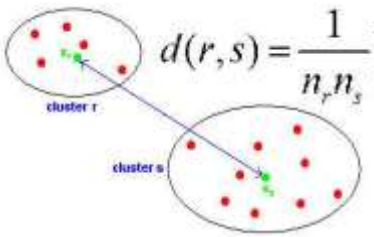
- فاصله بين دسته‌ها بر حسب دورترين فاصله ممكنه بين عناصر آنها محاسبه مي‌شود.



$$d(r,s) = \max\left(dist\left(x_{ri}, x_{sj}\right)\right)$$

Average linkage

- فاصله بين دو دسته مساوي مقادير متوسط كليه فاصله‌هاي ممكنه بين عناصر دو دسته است.

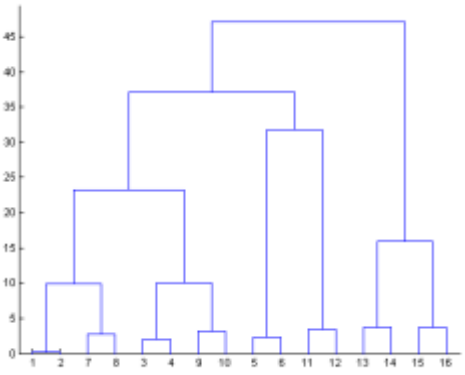


$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist\left(x_{ri}, x_{sj}\right)$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

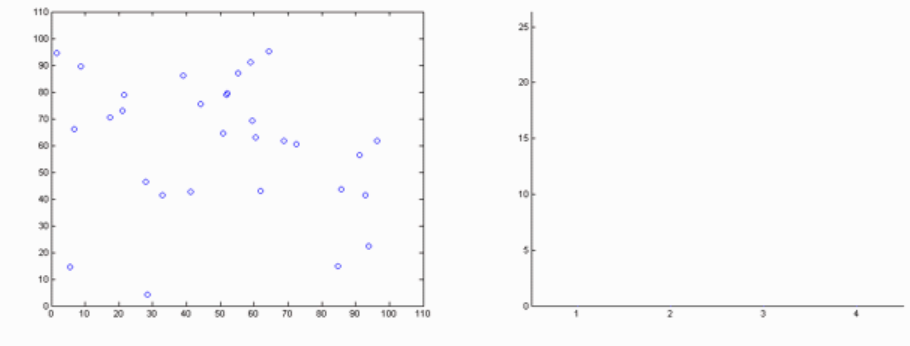
خوشه بندی سلسله مراتبی

- *به نمودار درختي تشكيل شده گفته* ***dendogram*** *مي‌شود .*



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

خوشه بندی سلسله مراتبی



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

مقایسه خوشه‌بندی سلسله‌مراتبی و غیر سلسله‌مراتبی

- روشهاي خوشه بندي غير سلسله مراتبي معمولا سريعتر عمل مي كنند ولي نياز به يكسري تصميم گيري از طرف تحليل گر و استفاده كننده دارد.
- انتخاب تعداد خوشه ها از اين گونه تصميم گيريها مي باشد.
- در اين گونه روشها معمولا يكسري خوشه هاي اوليه ايجاد شده و سپس در مراحل بعدي بهبود صورت مي گيرد.
- از آنجاييكه در اين روشها مناسب بودن خوشه ها به تعداد خوشه ها و يا حتي خوشه هاي اوليه بستگي دارد اين روشها كمتر از روشهاي خوشه بندي سلسله مراتبي انجام مي شوند.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی