مباحث تکمیلی در خوشه بندی

تعیینK – تعداد خوشه ها

- اگر تمام متغيرها كاملا مستقل باشند هيچ خوشه اي ايجاد نمي شود (تمام فضا بصورت تصادفي با نقاط داده ها پر مي شود) بر عكس اگر تمام متغيرها وابسته باشند آنگاه تمام داده ها تشكيل يك خوشه مي دهند.
- در شرايط بين استقلال و وابستگي كامل ما نمي دانيم كه واقعا چند خوشه وجود دارد.
- معمولا در انتخاب مقدار K نقش شخص تحليل گر بسيار بيشتر از كامپيوتر مي‌باشد.
- براي همين با توجه به كاربردهاي متفاوت روشهاي خوشه بندي و در جاهاي متفاوت ممكن است به تعداد بيشتر يا كمتري از خوشه ها نياز باشد.

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تعيينK – تعداد خوشه ها (ادامه)

- در بسياري از موارد با يك مقدار K خوشه بندي را انجام داده ، نتايج را بررسي مي كنند و دوباره به سراغ يك K ديگر مي روند.
- بعد از هر بار انجام اين كار قدرت و ارزش نتايج را بوسيله اندازه گيري ميزان متوسط فواصل در داخل خوشه ها و ميزان متوسط فواصل بين مراكز خوشه ها و بسياري روشهاي ديگر بررسي مي‌كنند.

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

تعيينK – تعداد خوشه ها (ادامه)

- تعداد خوشه ها را مي‌توان به اندازه تعداد داده‌هاي داخل پايگاه داده در نظر گرفت كه در اين حالت داده هاي هر خوشه دقيقا به هم متشابه هستند و داده هاي خوشه هاي متفاوت هم با هم تفاوت دارند (در هر خوشه تنها يك داده وجود دارد).
- ولي در اين حالت ديگر استفاده‌اي از خوشه بندي نمي توان كرد زيرا علت استفاده از خوشه بندي يافتن الگو هاي مناسب در پايگاه داده و تلخيص آنها بمنظور فهم بهتر پايگاه داده ها مي باشد.
- پس تعداد خوشه ها از تعداد داده ها كمتر مي باشد .
- اگر تعداد خوشه ها يكي در نظر گرفته شود تمام داده ها در واقع در آن خوشه قرار دارند و اطلاعات مفيدي كسب نمي شود .

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

تعيينK – تعداد خوشه ها (ادامه)

مزيت خوشه بندي سلسله مراتبي اين است كه به استفاده كننده و تحليل گر اجازه مي دهد كه از بين حالات مختلف يك عدد براي تعداد خوشه ها انتخاب نمايد.

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

ارزیابی کیفیت خوشه‌بندی ۱

- **دو سؤال اصلي**:
  - چه تعداد خوشه مناسب است؟
  - كدام تخصيص بهترين تخصيص نمونه‌ها به خوشه‌ها است؟

➢ **دو مشكل عمده**:

- تنظيمات اوليه پارامترهاي الگوريتم‌ها
- مقدار دهي اوليه الگوريتم‌هاي خوشه‌بندي

سميه عليزاده هيات علمي دانشكده صنايع
دانشگاه خواجه نصير طوسي

ارزیابی کیفیت خوشه‌بندی ۲

- **کمترین شباهت بین خوشه‌اي و بیشترین شباهت درون‌خوشه‌اي**

شاخص دیویس- بولدین

- **داشتن خوشه‌هاي متراکم با مرزهاي مشخص**

شاخص دان

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شاخص دان (داشتن خوشه‌هاي متراکم با مرزهاي مشخص)

- **در شاخص دان دو معیار در نظر گرفته میشود:**
- **حداکثر فاصله درون خوشه ای Intercluster**
- **حداقل فاصله بیرون خوشه ای Intracluster**

$$D = \frac{d_{min}}{d_{max}},$$

- **صورت کسر حداقل فاصله بین دو عنصر از دو خوشه متفاوت را بررسی می کند.**
- **مخرج کسر بیشترین فاصله بین دو عنصر درون یک خوشه را نشان میدهد.**
- **D در بازه [0.∞] است و هر چه بیشتر باشد بهتر است.**

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شاخص دان (داشتن خوشه‌هاي متراكم با مرزهاي مشخص)

$$D_{nc} = \min_{t=1,..,nc} \left\{ \min_{j=t+1,..,nc} \left(\frac{d(c_t, c_j)}{\max_{k=1,..,nc} diam(c_k)} \right) \right\}$$

$$d(c_t, c_j) = \min_{x \in c_t, y \in c_j} d(\mathbf{x}, \mathbf{y}),$$

$$diam(C) = \max_{x,y \in C} d(x, y)$$

$$\max_{nc=1,2,..n} \{D_{nc}\}$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شاخص دیویس–بولدین 1

- شاخص شباهت بين خوشه‌ها و بر اساس شاخص پراكندگي و شاخص تفاوت بين دو خوشه تعريف مي‌شود.

$$s_i = \frac{1}{C_i} \sum_{x \in C_i} \{\|x - z_i\|\},$$

$$d_{ij} = \|z_i - z_j\|,$$

$$R_{ij} = \frac{(s_i + s_j)}{d_{ij}}$$

$R_{ij} \geq 0.$

$R_{ij} = R_{ji}.$

if $s_i = 0$ *and* $s_j = 0$, *then* $R_{ij} = 0.$

if $s_j > s_k$ *and* $d_{ij} = d_{ik}$ *then* $R_{ij} > R_{ik}.$

if $s_j = s_k$ *and* $d_{ij} < d_{ik}$ *then* $R_{ij} > R_{ik}.$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

شاخص دیویس–بولدین ۲

- DB_{n_c} میانگین شباهت بین هر خوشه و شبیه‌ترین خوشه به آن می‌باشد.

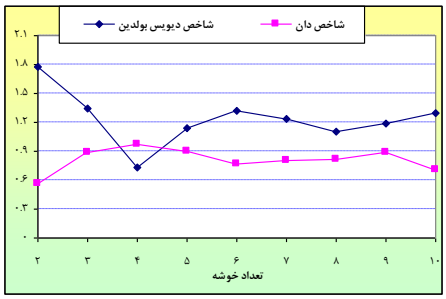
$$R_i = \max_{i=1,\ldots,n_c, i\neq j} R_{ij} ,$$

$$DB_{nc} = \frac{1}{n_c}\sum_{i=1}^{n_c} R_i ,$$

$$\frac{1}{C}\sum_{k=1}^{C}\max_{l\neq k}\left\{\frac{S_c(Q_k)+S_c(Q_l)}{d_{ce}(Q_k,Q_l)}\right\}$$

سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی

تعداد خوشه	شاخص دیویس- بولدین	شاخص دان
2	1.77	563/0
3	1.34	889/0
4	0.37	**975/0**
5	1.14	894/0
6	1.32	766/0
7	1.23	801/0
8	1.1	822/0
9	1.19	787/0
10	1.3	709/0



سمیه علیزاده هیات علمی دانشکده صنایع
دانشگاه خواجه نصیر طوسی