"بسمه تعالی"

پیاده سازی دسته بندی سلسله مراتبی به روش BIRCH

درس : داده کاوی

استاد : جناب آقای دکتر احمدی

تهیه کننده : فرخنده زینالی آق قلعه

شماره دانشجویی : ۹۶۱۱۲۷۴

بهار ۹۸

خوشه بندی (clustering) سلسله مراتبی بر دو نوع می باشد یکی تجمیعی و دیگری تقسیمی که در اینجا یکی از روش های تجمیعی به نام BIRCH را مورد استفاده قرار می دهیم. همچنین این نوع الگوریتم قابلیت بهبود یافتن نیز دارد که در الگوریتم AGNES امکانپذیر نبود.

در روش های تجمیعی ابتدا هر نمونه در یک خوشه قرار گرفته سپس با ادغام خوشه ها بر اساس معیار مشخصی خوشه ها کمتر شده تا در نهایت همه نمونه ها در یک خوشه قرار گیرند.

در پیاده سازی حاضر دیتاستی که باید مورد بررسی قرار گیرد ابتدا می بایست جهت استفاده در برنامه تنظیم و تصحیح شود. مراحل تنظیم دیتاست به شرح زیر انجام می شود:

- از آنجایی که برنامه برای داده های دو بعدی نوشته شده است لذا از میان خصوصیات نمونه ها در فایل دیتاست خصوصیت های ForkVA و ForkW را که اعدادی اعشاری از نوع real هستند را نگه می داریم.
- انواع نونه ها که شامل موارد

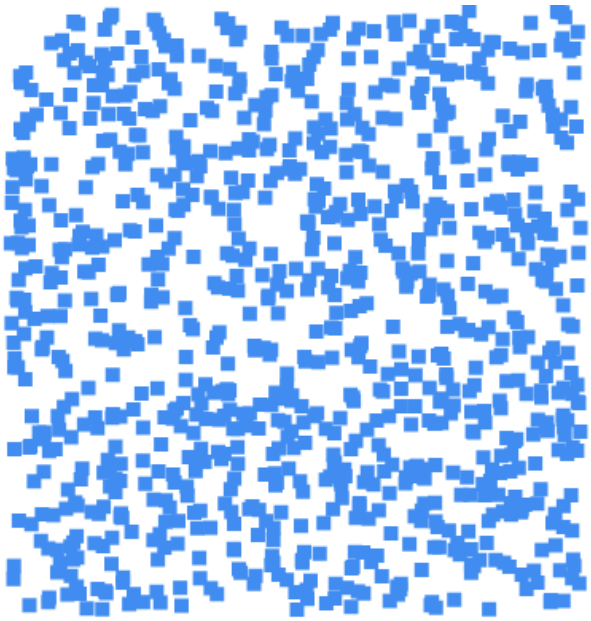
  'Bank' , 'AutomobileIndustry' , 'BpoIndustry' , 'CementIndustry' , 'Farmers1' , 'Farmers2' , 'HealthCareResources' , 'TextileIndustry' , 'PoultryIndustry' , 'Residential(individual)' , 'Residential(Apartments)' , 'FoodIndustry' , 'ChemicalIndustry' , 'Handlooms' , 'FertilizerIndustry' ,'Hostel' ,'Hospital' , 'Supermarket', 'Theatre' , 'University'

  می باشد را به ترتیب از شماره 0 تا 19 جایگزین می کنیم.
- از آنجایی که تعداد نمونه ها بسیار زیاد است لذا به صورت رندوم تعدادی از آنها را انتخاب کرده و بقیه را حذف می کنیم.

پس از تنظیم دیتاست می توانیم مراحل خوشه بندی را آغاز کنیم.

۱- معرفی فایل دیتاست به برنامه

فایل دیتا ست را به برنامه معرفی می کنیم. سپس برنامه با خواندن دیتا ست جدول نمونه ها را پر کرده و همچنین روی دستگاه مختصات نیز آنها را نمایش می دهد.



۲- عملیات خوشه بندی

در این مرحله عملیات خوشه بندی انجام می شود همانطور می دانید خوشه بندی به روش BIRCH بر اساس سلسله مراتب CF یا Clustering Feature که به صورت زیر تعریف می شود:

CF = (*N*, *LS*, *SS*)

***N* = |C|** **تعداد نقاط داده**

$$LS = \sum_{i=1}^{N} \overrightarrow{X_i}$$

جمع خطی n داده

$$SS = \sum_{i=1}^{N} \overrightarrow{X_i}^2$$

جمه مربع n داده

۳- اطلاعات آماری خوشه ها

با استفاده از محاسبات آماری چند ویژگی می توان اطلاعات ادغام هر خوشه را جهت مراحل بعدی انجام خوشه بندی سلسله مراتبی را بدست آورد. ویژگی های مذکور به شرح زیر است:

گرانیگاه خوشه

$$\overrightarrow{X_0} = \frac{\sum_{i=1}^{N} \overrightarrow{X_i}}{N}$$

شعاع خوشه

$$R = \left(\frac{\sum_{i=1}^{N} (\overrightarrow{X_i} - \overrightarrow{X_0})^2}{N}\right)^{1/2}$$

قطر خوشه

$$D = \left(\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} (\overrightarrow{X_i} - \overrightarrow{X_j})^2}{N(N-1)}\right)^{1/2}$$

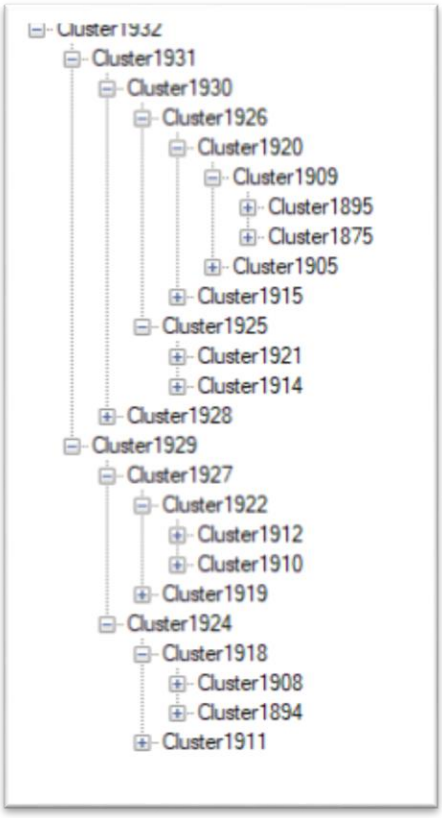
۴- ادغام خوشه ها

روش ادغام خوشه های ۱ با خصوصیت خوشه ای CF_1 و ۲ با خصوصیت خوشه ای CF_2 که در روش تجمیعی مورد استفاده قرار می گیرد از فرمول زیر استفاده می کند:

$$CF = CF_1 + CF_2$$

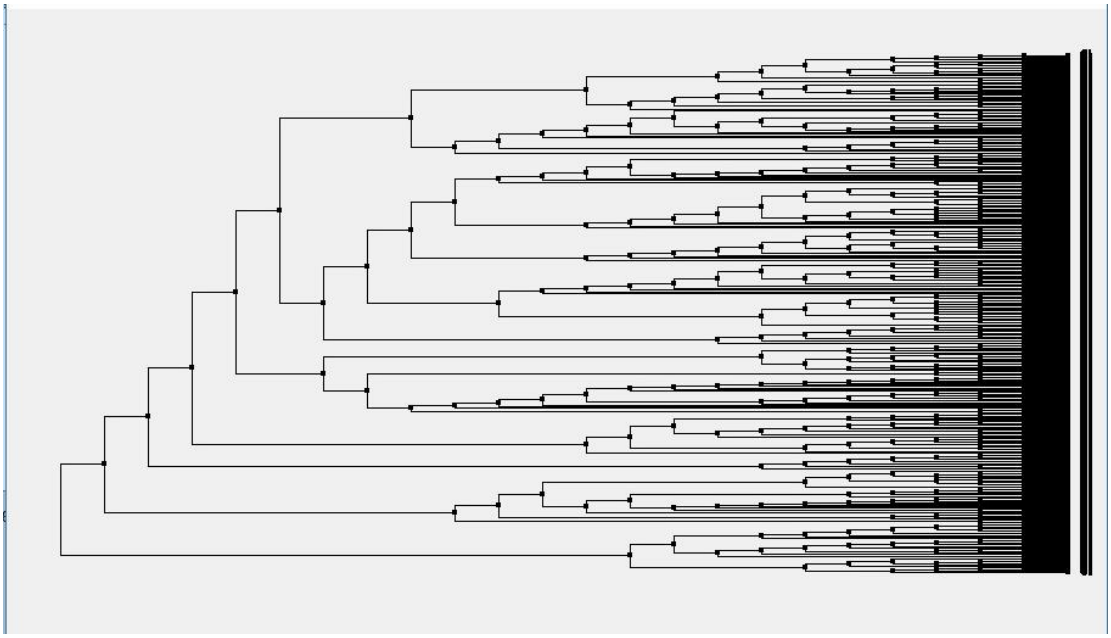
$$= (N_1 + N_2, \overrightarrow{LS_1} + \overrightarrow{LS_2}, SS_1 + SS_2)$$

بخشی از مراحل ادغام در زیر قابل مشاهده است:



۵- مشاهده دندروگرام

همانطور که می دانیم جهت نمایش خوشه بندی های سلسله مراتبی می توان از دندروگرام استفاده نمود که نمایش مناسبی برای خوشه بندی محسوب می شود پس از انجام عملیات خوشه بندی دندروگرام خوشه ها نیز قابل مشاهده خواهد بود.



۶- مشاهده خوشه بندی نمونه ها

پس از انجام عملیات خوشه بندی نمونه ها در هر یک خوشه قرار گرفته می شوند که می توان آنها را در دستگاه مختصات مشاهده و اعتبارسنجی نمود. در شکل های زیر نمونه ها را در ۲۰ خوشه ملاحظه می کنید.