به نام خدا

پروژه‌ی سوم داده کاوی

پیاده سازی FP Growth و یکی از روش های Apriori یا Eclat

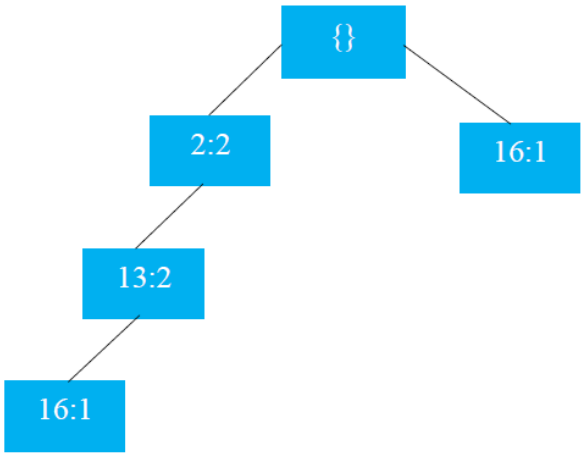
در این تمرین قصد داریم بخشی از مفهوم FP_Growth و یکی از روش های Apriori یا Eclat را بر روی یک مجموعه داده پیاده سازی کنیم. مجموعه داده مورد نظر در این تمرین، مجموعه داده ای است که در تمرین اول نیز مورد استفاده قرار گرفت. به یاد داریم که در مجموعه داده مذکور تعداد ۱۰.۰۰۰ رکورد وجود داشت که به صورت کلی شامل ۲۰ نوع آیتم است و هر کدام از این رکورد ها شامل تعدادی آیتم بودند.

در این پیاده سازی هر کدام از رکورد ها به منزله ی یک تراکنش هستند و ما قصد داریم درخت مربوطه یا FP_Treeرا برای تمامی این تراکنش ها تشکیل دهیم .توجه داریم که به صورت کلی الگوریتم FP_Growth شامل ساخت FP_Tree ، ساخت Conditional Pattern و همچنین پیدا کردن Frequent Patternsمی باشد اما در این تمرین برای ساده کردن پیاده سازی تنها به ساختFP_Tree برای مجموعه داده بسنده می کنیم.

به صورت کلی با روال FP_Growth آشنا هستیم اما برای تفهیم بهتر مطلب برای پیاده سازی یک مثال ساده را بررسی می نماییم .دو رکورد یا تراکنش زیر که از مجموعه داده انتخاب شده اند را در نظر بگیرید:

2 2 6 13 13 16
7 9 11 16 17

فرض نمایید که Min_Sup را در این مثال برابر با ۲ درنظر بگیریم ، در این صورت اگر بر روی این دو رکورد به تنهایی عملیات مربوطه را اجرا نماییم و FP_Tree آن را رسم نماییم ، شکل نهایی به این صورت خواهد شد:



از آنجایی که در مجموعه داده مربوطه دارای ۲۰ نوع آیتم هستیم ، ساخت FP_Tree و دو روش دیگر برای آن ها ممکن است کمی دشوار و سنگین باشد و درخت مربوطه بسیار بزرگ شود .از این رو دانشجویان عزیز کافی است که برای پیاده سازی این تمرین تنها آیتم های شماره صفر تا(۹ در مجموع ۱۰ نوع آیتم)را در نظر بگیرند . برای انجام این کار دانشجویان باید یک پیش پردازش بر روی مجموعه داده ی مربوطه انجام دهند و آیتم های ۱۰تا ۱۹را حذف نمایند.

نکته ی مهم در ارتباط با این تمرین این مسئله است که خروجی پیاده سازی باید به نحوی باشد که ساختار درخت مربوطه قابل مشاهده و تفسیر باشد .برای مثال قابل مشاهده و تفسیر باشد که کدام آیتم دارای بیشترین Frequencyاست(فرزندان اول ریشه نمایانگر این خصوصیت هستند) در آخر توجه نمایید که در این پیاده سازی شما باید Min_Sup مربوط به مسئله را تعیین نمایید.

پیاده سازی شما باید به گونه ای باشد که امکان تغییر Min_Sup و مشاهده نتیجه بر اساس تغییرات، وجود داشته باشد.

نکات مربوط به پیاده سازی و گزارش:

محدودیتی برای زبان پیاده سازی وجود ندارد و دانشجویان می توانند با زبان دلخواه پیاده سازی مربوطه را انجام دهند.

ساختار FP_Tree باید به صورت کامل پیاده سازی شود و در خروجی پیاده سازی به نحوی مشاهده شود.

توجه داشته باشید که یکی از روش های Apriori یاEclat را به دلخواه پیاده سازی کرده و خروجی باید مشاهده شود

تمرین شما باید شامل دو پوشه باشد به نام های Report و Souce Code باشد که در پوشه ی اول گزارش مربوطه و در پوشه ی دوم پیاده سازی انجام شده قرار داده می شود .فایل نهایی، فایل Zip شده ی این دو پوشه است که باید با فرمت زیر نام گذاری شود:

DataMining_Exc#3_Your Name_Your Student ID