Семинарски рад: Кластеровање високодимензионих података у C++

Увод

Високодимензиони подаци о генетским експресијама представљају посебно изазовну врсту података у анализи биолошких и медицинских истраживања.
Карактеристични су по томе што садрже велики број атрибута (гена), док је број узорака често релативно мањи.
То доводи до тога да је потребно пажљиво приступити њиховом предобрађивању и избору метода кластеровања које могу да открију скривене структуре у подацима.

Циљ нашег рада био је да испитамо примену и понашање различитих метода кластеровања на великом скупу података, са посебним фокусом на алгоритме OPTICS и CLARA, као и да прикажемо њихову имплементацију у програмском језику C++.

———

Опис скупа података

Добијени скуп података има димензију од отпрлике 12000 × 30000, што значи да садржи 12.000 атрибута и 30.000 инстанци који представљају вредности експресије различитих гена.
Због овако велике димензионалности, било је неопходно урадити одговарајуће предобрађивање како би подаци били употребљиви за кластеровање.

———

Предобрада података

Први корак у обради био је уклањање атрибута код којих су све вредности једнаке нули.
Такви атрибути не садрже никакву информацију и само повећавају димензионалност без користи.
Након тога, из скупа су издвојени 2000 атрибута са највећом варијансом.
Варијанса је коришћена као мера релевантности, јер атрибути са већом варијансом обично носе више информације и боље разликују узорке.
На редукованом скупу података затим је примењена PCA анализа (Principal Component Analysis) ради даљег смањења димензионалности и елиминације корелација између атрибута.
После PCA декомпозиције, подаци су редуковани на 50 компоненти, што је обезбедило знатно мањи и компактнији скуп погодан за примену алгоритама кластеровања.

———

Кластеровање OPTICS алгоритмом

Први алгоритам који је примењен био је OPTICS (Ordering Points To Identify the Clustering Structure). Овај алгоритам је сличан DBSCAN-

у, али је погоднији за откривање кластера различите густине.

OPTICS је омогућио да се добије хијерархијска структура кластера и да се лакше уоче границе између група података. У нашем случају, коришћени су следећи параметри:

- Epsilon (ε) = 1.45 – дефинише максималну удаљеност у којој се траже суседи;
- minNumberOfNeighbors = 10 – минималан број суседа потребан да би нека тачка била „core point“;
- clusterThreshold = 0.6 – праг који одређује границу за издвајање кластера из OPTICS дијаграма.

Применом ових параметара, добијено је 3 кластера различитих величина. Алгоритам је имплементиран у C++ језику без коришћења спољних библиотека, са функцијама за израчунавање растојања и претрагу суседа.

———

Кластеровање CLARA алгоритмом

Други алгоритам који смо применили био је CLARA (Clustering Large Applications). То је побољшана верзија K-medoids алгоритма, дизајнирана да ради на великим скуповима података тако што више пута узима узорке различитих величина и на њима врши кластеровање.
У овом случају смо, ради упоредивости са OPTICS резултатима, такође подесили да број кластера буде 3.
CLARA смо извршавали са различитим величинама узорака, на сликама су коришћени узорци величине 800, како бисмо испитали како величина узорка утиче на стабилност резултата. Уочено је да већи узорци дају стабилније резултате, али и да значајно продужавају време извршавања.

———

Визуелизација резултата

Да бисмо боље разумели добијене резултате, извршена је визуелизација кластера у 2D и 3D простору.
За те приказе коришћене су прве две, односно три компоненте добијене из PCA анализе.
Визуелизација је имплементирана тако сто смо у коду направили HTML и SVG фајлове који могу да се покрену у браузеру.

Визуелизација је омогућила да се лакше уоче груписања и односи између кластера.
Уочено је да OPTICS чешће формира неправилне, густо повезане кластере, док CLARA даје јасније раздвојене, али понекад мање прецизне групе.

———

Закључак

У овом раду приказали смо поступак анализе високодимензионих података о генетским експресијама – од предобраде и редукције димензионалности, до примене два различита алгоритма кластеровања и визуелизације резултата.
Алгоритам OPTICS показао се погодним за откривање сложених структура и кластера различите густине, али је захтевао више времена због обраде великог броја суседних тачака.
Са друге стране, CLARA је био знатно бржи и једноставнији за примену, али осетљив на избор узорка.
Као потенцијално проширење овог рада, могло би се испитати понашање других алгоритама као што су DBSCAN, BIRCH или CURE, као и употреба различитих метрика растојања које би могле боље описати сличност између гена.

Лазар Цвијић
Стојан Костић