Курсова работа по извличане на
закономерности от данни на
Ивайло Странджев
и Борис Странджев

Агломеративна йерархична клъстеризация

Задание:

Да се реализира алгоритъм за агломеративна йерархична клъстеризация с подходяща визуализация. Алгоритъмът трябва да е така реализиран че да дава възможност за експериментиране с различни подходи. Да се реализират подходящи мерки за оценка на алгоритъмът за клъстеризация.

Описание на алгоритъмът

Алгоритъмът за йерархична клъстеризация се състои от последователно сливане на клъстери до достигане на един единствен клъстер. Началното състояние е: всеки обект сам образува клъстер и на всяка стъпка сливаме двата клъстера, които са най-близо спрямо някаква мярка. В нашата реализация използваме няколко мерки за разстояние между два клъстера:

- Разстояние между най-близки обекти: за разстояние между два клъстера считаме най-малкото разстояние между двойка обекти като единия принадлежи на единия клъстер,а другия - на втория клъстер.
- Разстояние между най-далечни обекти: за разстояние между два клъстера считаме най-голямото разстояние между двойка обекти като единия принадлежи на единия клъстер,а другия - на втория клъстер.
- Разстояние между центроиди: за всеки клъстер изчисляваме неговия центроид - обект, чийто координати са средно аритметично на координатите на обектите в дадения клъстер. За разстояние между клъстерите използваме разстоянието между центроидите.
- Средно разстояние – за разстояние между двата клъстера се ползва средното аритметично на разстоянието между всички двойки обекти принадлежащи към двата клъстера.

Визуализация на алгоритъма

За визуализиране на отделните етапи от алгоритъма е използван код написан на C++, който използва OpenGL. Визуализацията е изцяло триизмерна и дава възможност за промяна на гледната точка с цел по-голяма свобода на наблюдението. Поради естествени ограничения тази визуализация може да изобразява само случая когато обектите са до три измерни. Освен това отново с цел по-ясна и естествена визуализация системата работи само с непрекъснати атрибути(макар и на теория клъстеризиращата част да може да се справи с номинални такива с относително малки изменения).

Визуализацията показва последователните етапи от сливането на клъстерите като дава възможност за преход между различните стъпки от алгоритъмът. На всяка стъпка обектите от един и същи клъстер са представени от различни(по форма и/или цвят) фигури, а при сливането на два клъстера новополученият клъстер получава фигурата на един от двата клъстера, които са се слели в него.

Оценка на алгоритъма

Реализирани са два алгоритъма за оценка на точността на клъстеризация.

- Алгоритъм базиран на кофенетична мярка – ненаправляван алгоритъм за оценка на точността на клъстеризация, базиращ се на разстоянието на сливане на двойките клъстери.
- Алгоритъм базиран на F-мярка – направляван алгоритъм за оценка на точността на клъстеризация, който оценява различието между намерената от нашия алгоритъм клъстеризация и реално известно предварително разделяне по класове. F-мярката се базира на мерките прецизност и припомняне, които се изчисляват за всеки клас.

Примери

Заедно с алгоритъмът са приложени и няколко примера използвани за оценка и проверка на коректната работа на решението.

С цел проверка на направляваната оценка, генерирахме 5 групи от случайно разпределени точки в рамките на квадрати с фиксирани размери. По този начин се получават 5 класа с частично засичане и можем да оценим по описания метод клъстеризацията на получените точки.

Резултати

Поради голямата изчислителна сложност на описания алгоритъм(N*N*log(N) където N е броя обекти) той не е подходящ за прилагане на големи по обем данни. Поради тази причина в подобни ситуации е подходяща комбинация между алгоритъмът за йерархична клъстеризация и други клъстеризиращи алгоритми.

От друга страна йерархичната клъстеризация би могла да обясни защо два обекта са попаднали в един и същ клъстер. Друго предимство е че в зависимост от нуждите си потребителя може да използва клъстери на различно ниво от йерархията като по този начин регулира броя и „качеството” им.