

Iteratív algoritmusok folytonos változók diszkrét kategóriákba sorolásához

Bali Attila

II. évf. programtervező informatikus MSc

Témavezető: Dombi József

SZTE TTIK Számítógépes Algoritmusok és Mesterséges Intelligencia Tanszék

Számos területen nagy jelentőséggel bír az adatok csoportosítása, amikor egy bizonyos adathalmaz elemeit jól elkülönített osztályokba kell sorolnunk. Ilyen terület például a természetes nyelvfeldolgozás, amikor nyers adatokból kell kinyernünk az ember számára érthető és könnyen értelmezhető információkat. Az adattárházak tipikus problémája ez, amikor hatalmas mennyiségű numerikus adatunk van, csak az információ kinyerése nehézkes.

Hasonlóan fontos a diszkrét csoportok határainak megtalálása a közgazdasági számításokban valamint a többtényezős döntések kiértékelésében.

A jelen dolgozat témájához két algoritmust implementáltunk, mindkét algoritmus hasonlóan diszkrétizál, de más körülmények között. Az első algoritmus valós számok (A, B) intervallumának szakaszokra osztását végzi, amivel előre megadott számú osztályba tudjuk sorolni az elemeket. A kialakult klaszterek a teljes (A, B) intervallumot lefedik, miközben egymás között nem lehet átlapolás. A másik megvalósítás, akkor hasznos, ha nem tudunk nyilatkozni a bemeneti adatok granularitásáról. Ekkor az algoritmus határozza meg a szükséges klaszterek számát és a felbontást. Ebben az esetben az adatsoron a sűrűsödési pontok között lyukakat keresünk, amik alapján szétbontjuk az adatokat, majd iteratív módszerrel egyesítés és szétbontás műveletét végző meta-algoritmus végzi a megfelelő csoportosítást, amely során kialakulnak a klaszterek határai.

Az algoritmusok implementálása során fontos kritérium volt a sebesség, mivel a problémák, amiket az algoritmus kiszolgál, általában nagy bemeneti adatsoron dolgoznak. Mindkét algoritmus iteratív módszere révén már az első iteráció végén is képes közelítő eredményt szolgáltatni, ami a további iterációkban javulni fog.

A fent tárgyalt algoritmusok jelenleg 1 dimenziós adatsoron működnek. A témavezetőm által már publikált képletek könnyen igazodhatnak akár több dimenziós klaszterezés algoritmusaihoz is. Több dimenziós esetben jobb közelítést tudunk adni többtényezős döntések kiértékeléséhez. A klaszterek és lyukak egyesítésével és szétosztásával lehetőségünk nyílik konkrét alakzatok klaszterezésére is, ami hasznosnak bizonyulhat a képfeldolgozás, alakfelismerés területeken is.