# Szépe Tamás

**III. programtervező matematikus, IV. műszaki informatikus**
**SZTE TTK**

**Konzulens: Dr. Kocsor András**
**tudományos főmunkatárs**


**Az X-Means klaszterező eljárás finomítása**

Napjainkban egyre népszerűbbek a különböző adatbányászati eljárások, melyek közül kiemelt szerepe van a felügyelet nélküli klaszterező algoritmusoknak. Ilyen algoritmus az ismert K-Means továbbfejlesztésére alapuló X-Means, melyet Dan Pelleg 2000-ben publikált először (*Dan Pelleg and Andrew Moore: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, ICML00*).

Az algoritmus a teljes adathalmazra alkalmazza a K-Means klaszterező eljárást kis kezdeti k értékkel, majd a keletkezett klasztereket egyesével vizsgálja aszerint, hogy érdemes-e tovább bontani két kisebbre. Ennek eldöntésére az un. *Bayesian Information Criteria*-t (*BIC*) használja, amely a pontok logaritmikus valószínűsége alapján számítható. Az algoritmus egy intervallumban keresi a klaszterek számának legjobb becslését az előbbi módszer iteratív alkalmazásával, és eredményül megadja a megfelelő számú középpontot.

Sokat gyorsít a pontok *KD-tree*-vel történő ábrázolása, viszont az eljárás csak kis dimenzióban (max 10) működik hatékonyan, valamint a *BIC* kizárólagos használatával csak az adathalmaz „sűrűsödéseit” fogja klaszteréknek venni, míg a klaszterek alakjából nyerhető információt nem veszi figyelembe. Ennek kiküszöbölésére mutat lehetőséget ez a tanulmány, amiben ismertetek egy olyan előfeldolgozó algoritmust, ami az adatpontok egymástól való távolsága alapján különválasztható szegmensekre bontja a bemenetet. A teszteredményekből kiderül, hogy milyen körülmények között jár előnyökkel ennek a módszernek a használata. Továbbá definiálom az X-Means algoritmus egy újszerű magasabb dimenziójú adatok feldolgozására is képes módosított változatát, amely eljárás létjogosultságát szimulációs eredményekkel igazolom.