Dimenziószám csökkentés nagy méretű bioinformatikai adathalmazok előfeldolgozásánál

(Reducing the number of dimensions on large bioinformatical data sets)

Témavezető: Kiss Attila (Waters Research Center, attila_kiss@waters.com)

A nagyméretű adathalmazok előfeldolgozása az egyik nagyon aktívan kutatott területe a modern alkalmazott matematikának mivel ennek előnyei az iparban egyre nagyobb ismerettségnek örvend. Ezen előfeldolgozás egyik fontos területe a dimenziók számának csökkentése. Bioinformatikai adathalmazokon tipikusan sokkal több paraméterünk van, mint mérési pontunk így fontos, hogy megszűrjük ezeket a paramétereket, hogy később hatékony osztályozó algoritmusokat tudjunk készíteni hozzájuk. Valahogy úgy kell elképzelni ezt a feladatot, mint amikor szeretnénk az embereket besorolni osztályokba (pl. tökéletes, csodálatos, elbűvölő, bámulatos) amely besorolást nem igazán tudjuk, hogy mi alapján kellene megtennünk, viszont rengeteg adatunk van minden egyes emberről (név, kor, magasság, nem, szemszín, kedvenc étel…stb.) és ezekből szeretnénk kiszűrni a besorolásukhoz szükséges attríbútumokat, amelyek alapján már biztosan és gyorsan be tudjuk őket sorolni a megadott osztályokba. A nálunk végzett kutató munka során alapvetően olyan adathalmazokhoz fejleszthetnek a hallgatók új dimenzió csökkentő algoritmusokat, amelyeknél a paraméterek száma jóval nagyobb, mint a minták száma. Célunk a minél több jelentéssel bíró paraméterek megszűrése, kiválasztása konkrét bioinformatika feladatokhoz. A hallgató bepillantást nyerhet a bioinformatikai adathalmazok előfeldolgozásának világába és kipróbálhatja, mélyítheti algoritmus fejlesztő tudását is, miközben új algoritmusokat dolgozhat ki kutatóink segítségével.

A jelentkezővel szemben támasztott elvárások:
– Alapszintű programozási ismeretek MATLAB és C++ vagy Python nyelveken.
– Angol nyelv ismerete előny, de nem feltétel.
– Előny ha vannak adatbányászati alapismeretek, “feature selection” algoritmusok ismerete, analitikus szemlélet, esetleg mélytanuló hálózatok ismerete

Cikkek:
– Yamada et al.: High-Dimensional Feature Selection by Feature-Wise Kernelized Lasso (https://arxiv.org/pdf/1202.0515.pdf)
– Balog et al.: Identification of the Species of Origin for Meat Products by Rapid Evaporative Ionization Mass Spectrometry (https://pubs.acs.org/doi/10.1021/acs.jafc.6b01041)