Hatékony osztályozó algoritmusok fejlesztése és implementálása bioinformatikai alkalmazásokhoz

(Development and implementation of effective classificator algorithms on biomedical data sets)

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

A nagyméretű adathalmazok feldolgozása az egyik nagyon aktívan kutatott területe a modern alkalmazott matematikának mivel széleskörű alkalmazhatósága az iparban egyre nagyobb ismerettségnek örvend. Ezen feldolgozás egyik fontos területe a beérkező adatok kategóriákba sorolása, osztályozása. Valahogy úgy kell elképzelni ezt a feladatot, mint amikor valaki ad nekünk rengeteg csomagot, amelyekről tudunk sok paramétert (szín, méret, szag, halmazállapot, kedvenc szórakozási mód…stb.) és ezek alapján be kell sorolnunk ezeket a csomagokat automatikusan egy-egy osztályba (pl.: fontos, vicces, hasznos, érdekes). A probléma ott kezdődik, amikor a tulajdonságaik alapján nagyon nehéz eldönteni, hogy egy adott csomag az inkább vicces vagy inkább hasznos, mivel nagyon közel esik hasonló csomagokhoz, amelyek közül néhányról tudjuk, hogy vicces, néhányról meg azt, hogy hasznosak, viszont azt is tudjuk, hogy minden csomag pontosan egy osztályba illik bele. A nálunk végzett kutató munka során alapvetően olyan adathalmazokhoz fejleszthetnek a hallgatók osztályozó algoritmusokat, amelyeknél a paraméterek száma jóval nagyobb, mint a minták száma. Célunk minél gyorsabb és természetesen minél pontosabb algoritmusok fejlesztése konkrét bioinformatika feladatokhoz. A hallgató bepillantást nyerhet a bioinformatikai adathalmazok feldolgozásának 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, osztályozó algoritmusok ismerete, analitikus szemlélet, esetleg mélytanuló hálózatok ismerete

Cikkek:
– Peng et al.: An Introduction to Logistic Regression Analysis and Reporting (http://sta559s11.pbworks.com/w/file/fetch/37766848/IntroLogisticRegressionPengEducResearch.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)