



Budapesti Műszaki és Gazdaságtudományi Egyetem  
Méréstechnika és Információs Rendszerek Tanszék

# Kilógó értékek detektálása adatfolyamokon

Raszler Gergő Imre

BSc, mérnök informatikus

Konzulens: Salánki Ágnes

Napjainkban egyre fontosabbá válik a rendszerekben levő hibák, rendellenességek vagy csalások valós időben történő keresése, jellemzése. Ez nemcsak a kilógó értékek keresésének általános feladatát (*outlier detection*) foglalja magában, hiszen az időablakos feldolgozásból adódóan az algoritmusok többségénél feltételezése, miszerint minden megfigyelés mindegyik másikkal összehasonlítható, nem áll fenn. Ez indokolja olyan online algoritmusok kifejlesztését, amelyek az adatfolyamok feldolgozását lehetővé tévő teljesítménykényszereknek megfelelnek és az eredmények egyszerű mérnöki interpretációjára is lehetőséget biztosítanak.

A félév során megismerkedtem a jellemzően használt, kilógó értékeket kereső algoritmusokkal, mint például a DB (Distance Based) és a LOF (Local Outlier Factor) megközelítések.

Az adatfolyamokkal dolgozó algoritmusok közül kettő state-of-the-art algoritmust, a szomszédok távolságával operáló Exact-Storm [1], valamint a klaszter alapú KORM [2] algoritmust Pythonban leimplementáltam. Az algoritmusokat az Apache Storm [3] adatfolyam-feldolgozó környezetbe integráltam, a módszerek sikerességét mesterséges és valós adatokon is teszteltem.

Hivatkozások

[1] Exact-STORM: <http://dl.acm.org/citation.cfm?id=1321552>

[2] KORM: <http://adsabs.harvard.edu/abs/2010arXiv1002.4003D>

[3] Apache Storm: <https://storm.apache.org/>