PhD. Thesis-Short Abstract

Title: **Density-Based Mining Algorithms for Dynamic Data: An Incremental Approach**

Panthadeep Bhattacharjee
136101011
Thesis Supervisor: Dr. Pinaki Mitra
Department of CSE
IIT Guwahati

Short Abstract

Typically an algorithm designed for carrying data mining tasks is fed with a static set of input. The prominent bottlenecks with this class of algorithms may include redundant computation, high response time along with increased consumption of available resources. Given the importance of handling dynamic data, this thesis focuses on developing incremental mining algorithms particularly in the field of density-based clustering and outlier detection. This is because density-based algorithms display robustness in extracting clusters of varying granularity or filtering outliers from variable density sub-spaces. We propose incremental extensions to two density based clustering algorithms: MBSCAN, SNN-DBSCAN and an outlier detection algorithm KNNOD. The incremental extensions to MBSCAN namely *iMass* and KNNOD known as KAGO are approximate in nature facilitating single point insertions. While for SNN-DBSCAN, we propose exact incremental solutions: $BISDB_{add}$, $BISDB_{del}$ facilitating both addition and deletion of data in batch mode. The *iMass* algorithm achieved a maximum efficiency upto an order of 2.28 ($\approx$ 191 times) maintaining a mean clustering accuracy of 60.375%. For labeled data *iMass* had similar or improved NMI (Normalized Mutual Information), RI (Rand index) measure for some datasets. $BISDB_{add}$ and $BISDB_{del}$ achieved a maximum efficiency upto an order of 3 and 4 respectively. The set of clusters obtained were identical to SNN-DBSCAN. KAGO outperformed KNNOD by achieving efficiency upto an order of about 3.91 ($\approx$ 8304 times) over two intrusion detection datasets and a bidding data related to a search engine. Outliers' evaluation

on these datasets using RI and F1-score showed a mean improved accuracy of around 3.3% for KAGO as compared to KNNOD.