Title:
NEIGHBORHOOD DENSITY ESTIMATION USING SPACE-PARTITIONING BASED HASHING SCHEMES

Abstract:
Single cell messenger RNA sequencing (scRNA-seq) offers a view into transcriptional landscapes in complex tissues. Recent developments in droplet based transcriptomics platforms have made it possible to simultaneously screen hundreds of thousands of cells. It is advantageous to use large-scale single cell transcriptomics since it could lead to the discovery of a number of rare cell sub-populations. When the sample size reaches the order of hundreds of thousands, existing techniques to discover rare cells either scale unbearably slow or terminate altogether. We suggest the Finder of Rare Entities (FiRE), an algorithm that quickly assigns a rareness score to every individual expression profile under consideration. We show how FiRE scores can assist bioinformaticians in limiting the downstream analyses to only on a subset of expression profiles within ultra-large scRNA-seq data.

Anomaly detection methods differ in their time complexity, sensitivity to data dimensions, and their ability to detect local/global outliers. The proposed algorithm FiRE is a 'sketching' based linear-time algorithm for identifying global outliers. FiRE.1, an extended implementation of FiRE fares well on local outliers as well. We provide an extensive comparison with 18 state-of-the-art anomaly detection algorithms on a diverse collection of 1000 annotated datasets. Five different evaluation metrics have been employed. FiRE.1's performance was particularly remarkable on datasets featuring a large number of local outliers. In the sequel, we propose a new "outlierness" criterion to infer the local or global identity of outliers.

We propose Enhash, a fast ensemble learner that detects concept drift in a data stream. A stream may consist of abrupt, gradual, virtual, or recurring events, or a mixture of various types of drift. Enhash employs projection hash to insert an incoming sample. We show empirically that the proposed method has competitive performance to existing ensemble learners in much lesser time. Also, Enhash has moderate resource requirements. Experiments relevant to performance comparison were performed on 6 artificial and 4 real datasets consisting of various types of drifts.