NEIGHBORHOOD DENSITY ESTIMATION USING SPACE-PARTITIONING BASED HASHING SCHEMES

AASHI JINDAL

DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI

June 2023

NEIGHBORHOOD DENSITY ESTIMATION USING SPACE-PARTITIONING BASED HASHING SCHEMES

by

AASHI JINDAL

DEPARTMENT OF ELECTRICAL ENGINEERING

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

June 2023

Certificate

This is to certify that the thesis entitled **"Neighborhood Density Estimation Using Space-Partitioning Based Hashing Schemes"**, being submitted by **Aashi Jindal** for the award of the degree of **Doctor of Philosophy** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, is a record of bonafide work done by her under my supervision and guidance. The matter embodied in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma.

Dr. Jayadeva

Professor

Department of Electrical Engineering,

Indian Institute of Technology Delhi,

Hauz Khas, New Delhi - 110016,

INDIA.

Dr. Shyam Prabhakar

Associate Director

Laboratory of Systems Biology and Data Analytics,

Genome Institute of Singapore,

SINGAPORE

Dr. Debarka Sengupta

Associate Professor

Department of Computer Science & Engineering,

Department of Computational Biology,

Centre for Artificial Intelligence,

Indraprashta Institute of Information Technology,

Okhla Phase III, Delhi - 110020, India

(Adj.) Associate Professor

Institute of Health & Biomedical Innovation,

QUT, AUSTRALIA

Acknowledgements

I sincerely thank my supervisors Dr. Jayadeva, Dr. Debarka Sengupta and Dr. Shyam Prabhakar for their valuable feedback. I also thank other SRC members Dr. I.N. Kar, Dr. Shouri Chatterjee and Dr. Vivekanandan Perumal. I thank Prof. Suresh Chandra for his encouragement. I thank my colleagues Dr. Sumit Soman and Dr. Udit Kumar for their moral support. I thank department staff members Rakesh, Yatindra, Satish and Mukesh for always helping with the department work.

I thank my parents, Rakesh Jindal and Meena Jindal, for always believing in me and supporting me in my decisions. I thank my parents-in-law, Sanjiv Kumar Gupta and Snehlata Gupta for being always understanding. I thank my husband Prashant Gupta for always being there whenever I have needed him. He supported me both professionally and personally. My brother, Sahil Jindal always holds a special place in my life. My core family members, Shruti Singla, Praveen Singla, Shaffi Bindal, Mohit Bindal, Khushboo Mehrotra, Yuvan Singla, Tejas Bindal and Vanya Bindal are an important part of my life.

I also thank my current employers, Ashok Juneja and Shweta Singla for giving me the time to complete my pending research work when needed. I am thankful to my company, Applied Solar Technologies India Pvt. Ltd. and my colleagues, Subhdip Rakshit, Amod Kumar, Shubham Sinha, Pulkit Tyagi, Bikas Gupta, Sukrit Singh Negi, Abhishek Chaudhary and Anil Hansda.

The following have been the three main philosophies of my life, "There is an almighty somewhere looking after us and guiding us to the correct path." "Behind every successful woman is her parents, husband and family that supported her in her unconventional choices." Last, but not the least, "...all's well, that ends well."

(Aashi Jindal)

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.

सार

एकक कोशिका दूत-RNA (scRNA-seq) अनुक्रमण जटिल ऊतकों के लिप्यंतरण का एक दृश्य प्रस्तुत करता है। छोटी बूंदों पर आधारित अत्याधुनिक प्रतिलेख यंत्रों के माध्यम से लाखों एकक कोशिकाओं को एक साथ परखना संभव हो गया है। बड़े पैमाने पर एकक कोशिकाओं के प्रतिलेखों का उपयोग दुर्लभ कोशिकाओं को खोजने में किया जा सकता है। जब एकक कोशिकाओं के प्रतिदर्शों का आकड़ां लाखों में पहुँचता है तो उपलब्ध कलन विधिआँ असहनीय रूप से शिथिल हो जाती हैं तथा इनका उपयोग करके दुर्लभ कोशिकाओं की खोज करना लगभग असंभव हो जाता है। इस शोध प्रबंध में दुर्लभता के अंकमान की गड़ना के लिए एक कलन विधि, दुर्लभ इकाईओं का खोजक (FiRE) को प्रस्तुत किया गया है। यह कलन विधि तीव्र गति से प्रत्येक कोशिकाओं के प्रतिलेख को एक दुर्लभता का अंकमान प्रदान करती है। हम यह भी दिखते हैं FiRE अंकमान जैव सूचना वैज्ञानिकों को अत्यधिक विशाल scRNA-seq आंकडाकोशों में से केवल कुछ महत्वपूर्ण एकक कोशिकाओं के विश्लेषण करने में सहायता करता है।

विसंगति का पता लगाने के तरीके उनकी समय जटिलता, विवरण आयामों, तथा स्थानीय/वैश्विक रूप से विचित्र प्रतिदृश्यों की खोज करने की छमता में भिन्न होते है। प्रस्तावित कलन विधि FiRE वैश्विक रूप से विचित्र प्रतिदृश्यों को रेखीय अवधि में अवगत कराती है तथा इसके लिए स्केचिंग, एक हैशिंग पे आधारिक कलन विधि का उपयोग करती है। FiRE का विस्तारित कार्यान्वयन FiRE.1 स्थानीय रूप से विचित्र प्रतिदृश्यों की खोज करने में सक्षम है। इस शोध प्रभंद में हम १००० मापदंड आंकड़ा संचयों के विविध संग्रह पे १८ अत्याधुनिक विसंगति का पता लगाने वाली कलन विधिओं की व्यापक तुलना प्रदान करते हैं। इस तुलना के लिए ५ भिन्न प्रक्कर की तुलना विधिओं को नियोजित किया गया है। बड़ी संख्या में स्थानीय रूप से विचित्र प्रतिदृशों से परिपूर्ण आंकड़ा संकलनो पर FiRE.1 का प्रदर्शन उल्लेखनीय है। अगली कड़ी में हम स्थानीय और वैश्विक रूप से विचित्र प्रतिदृश्यों का अनुमान लगाने के लिए एक "बाह्यता" मानदंड प्रस्तावित करते हैं।

इस शोध प्रबंध के अंतिम भाग में हम एक तीव्र समवेत कलन विधि, Enhash, को प्रस्तुत करते हैं जो आंकड़ों की धारा में से अवधारणा के बहाव का पता लगाने में सक्षम है। एक आंकड़ों की धारा में आकस्मिक, क्रमिक, आभासी, या आवर्ती घटनाये, या इनके मिश्रण उपास्थिक हो सकते हैं। Enhash आने वाले प्रतिदर्शियों को सम्मलित करने के लिए प्रक्षेप हैश का उपयोग करता है। हम अनुभवजन्य रूप से दिखते हैं की प्रस्तावित कलन विधि का प्रदर्शन उपलब्ध समवेत आधारित कलन विधिओं की अपेक्षा प्रतिस्पर्धी है तथा प्रस्तावित कलन विधि यह प्रदर्शन अन्य विधिओं की तुलना में कम समय तथा मद्यम संसाधनों का उपयोग करके प्राप्त करती है। प्रस्तावित कलन विधि के प्रदर्शन की तुलना ६ कृत्रिम तथा ४ वास्तविक आंकड़ा संचयों पर की गई है। यह आंकड़ा संचय विभिन्न प्रकार के अवधारणाओं के मिश्रण से बने हैं।

Contents

*The work presented in this chapter has been published as a research paper titled "*Discovery of rare cells from voluminous single cell expression data*" in Nature Communications (2018).

†The work presented in this chapter has been published as a research paper titled "*Linear time identification of local and global outliers*" in Neurocomputing (2021).

[‡]The work presented in this chapter has been published as a research paper titled "*Enhash: A fast streaming algorithm for concept drift detection*" in ESANN proceedings (2021).

List of Figures

List of Tables

List of Abbreviations

P@n *Precision at n*

***k*NN** k^{th} Nearest Neighbor

***k*NNW** *k*NN-weight

AP *Average Precision*

AUC *Area Under the Curve*

FPR *False Positive Rate*

ROC AUC *Receiver Operator Characteristics Area Under the Curve*

ROC *Receiver Operator Characteristics*

TPR *True Positive Rate*

ABOF Angle-based outlier factor

ADWIN ADaptive WINdowing

AEE Additive Expert Ensemble

ARF Adaptive Random Forest

AWE Accuracy-Weighted Ensemble

COF Connectivity-based Outlier Factor

CTCs Circulating Tumor Cells

DB-outlier score Distance-based outlier score

DC Dendritic Cell

DCs Dendritic Cells

DE differentially expressed

EPCs Endothelial Progenitor Cells

ESC embryonic stem cell

ESCs embryonic stem cells

FACS Fluorescence-activated cell sorting

FastABOD Fast Angle-Based Outlier Detection

FiRE Finder of Rare Identities

FP False Positives

HBOS Histogram-based Outlier Score

INFLO INFLuenced Outlierness

KDEOS Kernel Density Estimation Outlier Scores

LB Leveraging Bagging

LDE Local Density Estimate

LDF Local Density Factor

LDOF Local Distance-based Outlier Factor

LIC Local Isolation Coefficient

LOF Local Outlier Factor

LoOP Local Outlier Probabilities

LSH Locality Sensitive Hashing

OB Online Bagging-ADWIN

ODIN Outlier detection using Indegree Number

OSMOTEB Online SMOTE Bagging

PBMC peripheral blood mononuclear cell

PBMCs peripheral blood mononuclear cells

RMS Root Mean Square

scRNA-seq single-cell RNA sequencing

SNV Single Nucleotide Variant

SOD Subspace Outlier Degree

t-SNE t-Distributed Stochastic Neighbor Embedding

TN True Negatives