# TECHNIQUES TO SPEED-UP COUNTING BASED DATA MINING ALGORITHMS ON GPUS

Amila De Silva

168062J

Degree of Master of Science

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

April 2019

# TECHNIQUES TO SPEED-UP COUNTING BASED DATA MINING ALGORITHMS ON GPUS

Amila De Silva

168062J

Thesis/Dissertation submitted in partial fulfillment of the requirements for the degree Master of Science in Computer Science and Engineering

Department of Computer Science & Engineering

University of Moratuwa

Sri Lanka

April 2019

# DECLARATION

I declare that this is my own work and this dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date:

The above candidate has carried out research for the Masters thesis/Dissertation under my supervision.

Signature of the Supervisor: Date:

# ACKNOWLEDGEMENTS

I am sincerely grateful for the advice and guidance of my supervisors Dr. Shehan Perera and Prof. Sanath Jayasena. Without their help and encouragement this project would not have been completed. I would like to thank them for taking time out of their busy schedule to be available anytime that was needed with help and advice.

I would also like to thank my progress review committee, Dr. Surangika Ranathunga and Dr. Lochandaka Ranathunga. Their valuable insights and guidance helped me immensely.

I would like to thank the entire staff of the Department of Computer Science and Engineering, Both academic and non-academic for all their help during the course of this work and for providing me with the resources necessary to conduct my research.

My sincere gratitude goes to Senate Research Grant for funding this study. This work was partially funded by the LK Domain Registry through Prof. V.K. Samaranayake top-up grant.

Finally, I would like to express my gratitude to my family and all my friends for their support.

# ABSTRACT


## Techniques to speed-up counting based Data Mining Algorithms on GPUs

Data Mining by its definition is meant to deal with large volumes of data. Ever growing volumes of Data and increasing demand for data driven decisions are placing new requirements on Data Mining algorithms. To respond to these demands Data Mining practitioners are focusing on improving speed and turnaround time without compromising accuracy.

Among different approaches in improving speed, one approach gaining increased attention is the use of GPUs. Ability of GPUs to perform parallel executions at a massive scale and inherently repetitive nature of Data Mining workloads make GPUs a better candidate in improving speed.

Another area getting increased attention is using Bitmaps for Data Mining algorithms. Bitmap representations have been abundantly used in analytical queries for their ability to represent data concisely and for being able to simplify processing.

A number of studies have been carried out which combine these two techniques to achieve greater performance improvements. But most of those studies are revolving around FIM based algorithms, processing of which naturally aligns with Bitmap representations.

In this study, we explore the ability of using Bitmap techniques on GPUs to speed up a class of Data Mining Algorithms. A Counting based Algorithm can be defined as an Algorithm which can be separated into to two distinct phases a pattern counting phase and a model building phase. We propose a framework based on Bitmap techniques, which speeds up these counting based algorithms on GPUs. The proposed framework uses both CPU and GPU for the algorithm execution, where the core computing is delegated to GPU. We implement two algorithms Naïve Bayes and Decision Trees, using the framework, both of which outperform CPU counterparts by several orders of magnitude.



**Keywords**: Data Mining; GPU; Classification; Bitmaps;BitSlices;Naïve Bayes;Decision Trees

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS

| Abbreviation | Description |
|---|---|
| ARM | Associate Rule Mining |
| FIM | Frequent Itemset Mining |
| GPGPU | General purpose Graphic Processing Units |
| IOT | Internet of Things |
| SIMD | Single Instruction Multiple Data |