ANALYSIS OF ALTERNATIVE GRAPHICAL REPRESENTATION FOR THE SELF-ORGANIZING MAPPING OF THE SUPERSYMMETRY DATASET

BY

NU'MAN BIN BADRUD'DIN

A thesis submitted in fulfilment of the requirement for the degree of Master of Science (Computational and Theoretical Sciences)

Kulliyyah of Science
International Islamic University Malaysia

JULY 2021

ABSTRACT

High energy physics (HEP) simulation and experimentation data are often high dimensional containing high number of features. A beyond standard model (BSM) dataset that is the supersymmetry (SUSY) event simulation dataset was clustered using self-organising map (SOM) algorithm. SOM clustering is one of the better methods to cluster high dimensional data. To verify the existence of the SUSY event in the clustered dataset, it was visualised through several different methods which are the U-matrix, principal component analysis (PCA) and spectral graph theory. U-matrix is the default representation of SOM that visualises the distance between SOM neurons. PCA reduces the dimensionality of the dataset to only 2-D and 3-D considering only the principal components. Spectral graph connects all the neurons together as a network but the implementation was limited by computational resources due to connecting all the neurons of the high dimensional data requires much more intense computational power. While both U-matrix and PCA are successful in visualising cluster(s) in digit datasets, U-matrix was unsuccessful in showing cluster for the SUSY dataset. PCA on the other hand manages to display cluster existence in the SUSY dataset. This may suggest that U-matrix is limited to a certain number of dimensions and PCA might be a better option for cluster existence verification. Further research needs to be done to probe into the potential of dimensionality reduction of clustered HEP data. The visualisation of cluster existence hints to the potential of the algorithm to be used on actual experimentation dataset.

ملخص البحث

إن بيانات المحاكاة والتجريب في مجال فيزياء الطاقة العالية (HEP)، غالبا ما تكون عالية الأبعاد وتحتوي على كمية عالية من السمات. مجموعة بيانات خارج النموذج القياسي (BSM) التي هي قاعدة بيانات المحاكاة أحداث التناظر الفائق (SUSY) تم تجميعها باستخدام خريطة ذاتية التنظيم (SOM). وهي إحدى أفضل الطرق لتجميع البيانات عالية الأبعاد. للتحقق من وجود حدث التناظر الفائق (SUSY) في تلك البيانات، تم تصويره من خلال عدة طرق منها طريقة مصفوفة المسافة الموحدة (U-matrix) وتحليل المكونات الرئيسية (PCA) ونظرية الرسم البياني الطيفيّ (spectral graph theory). ومصفوفة المسافة الموحدة (U-matrix) هي التمثيل الافتراضي للخريطة ذاتية التنظيم (SOM) التي تصور الخلايا العصبية للخريطة. وتحليل المكونات الرئيسية (PCA) يقلل الأبعاد في البيانات إلى البعد الثاني (2D) والبعد الثالث (3D) بالنظر إلى المكونات الرئيسية فقط. والرسم البياني الطيفيّ يربط ويتصل كل الخلايا العصبية معا كشبكة ولكن كان تنفيذه محدودا بسبب الربط بين جميع الخلايا العصبية للبيانات عالية الأبعاد تتطلب قوة حسابية عالية الكثافة. في حين أن كلا من مصفوفة المسافة الموحدة (U-matrix) وتحليل المكونات الرئيسية (PCA) ناجحتان في تصوير مجموعات في البيانات الرقمية، فمصفوفة المسافة الموحدة-(U-matrix) لم تنجح في إظهار مجموعة في بيانات التناظر الفائق (SUSY). ومن ناحية أخرى فإن تحليل المكونات الرئيسية (PCA) يمكن أن يبين وجود المجموعات التناظر الفايق (SUSY). ومن هذا يمكن للمصفوفة المسافة الموحدة (U-matrix) أن تقتصر على عدد معين من الأبعاد ويكون تحليل المكونات الرئيسية (PCA) هو الخيار الأفضل للتحقق من وجود المجموعات. إن نتائج هذا البحث تشير إلى أنه يلزم إجراء مزيد من البحوث للنظر في امكانية تخفيض الأبعاد في البيانات المجمعة لفزياء الطاقة العالية (HEP). وامكانية نجاح تصوير وجود المجمعات في بيانات المحاكاة يشير إلى أن الخوارزمية يمكن أن يستخدم في بيانات التجارب الفعلية.

Reviewed and approved by
Assoc. Prof. Dr. Ibrahim Shogar
Kulliyyah of Science, IIUM-Kuantan

APPROVAL PAGE

I certify that I have supervised and read this study and that in my opinion, it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computational and Theoretical Sciences).

…………………………………..
Mohd. Adli bin Md. Ali
Supervisor

…………………………………..
Mohd Hirzie bin Mohd Rodzhan
Co-Supervisor

I certify that I have read this study and that in my opinion it conforms to acceptable standards of scholarly presentation and is fully adequate, in scope and quality, as a thesis for the degree of Master of Science (Computational and Theoretical Sciences).

…………………………………..
Azni binti Abdul Aziz
Internal Examiner

…………………………………..
Imran bin Yusuff
External Examiner

This thesis was submitted to the Department of Computational and Theoretical Sciences and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computational and Theoretical Sciences).

…………………………………..
Nurul Farahain binti Mohammad
Head, Department of
Computational and Theoretical
Sciences

This thesis was submitted to the Kulliyyah of Science and is accepted as a fulfilment of the requirement for the degree of Master of Science (Computational and Theoretical Sciences).

…………………………………..
Shahbudin bin Saad
Dean, Kulliyyah of Science

DECLARATION

I hereby declare that this thesis is the result of my own investigations, except where otherwise stated. I also declare that it has not been previously or concurrently submitted as a whole for any other degrees at IIUM or other institutions.

Nu'man bin Badrud'din

Signature .... Date 18 / 7 / 2021 ....

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA

DECLARATION OF COPYRIGHT AND AFFIRMATION OF FAIR USE OF UNPUBLISHED RESEARCH

ANALYSIS OF ALTERNATIVE GRAPHICAL REPRESENTATION FOR THE SELF-ORGANIZING MAPPING OF THE SUPERSYMMETRY DATASET

I declare that the copyright holders of this thesis are jointly owned by the student and IIUM.

No part of this unpublished research may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise without prior written permission of the copyright holder except as provided below

1. Any material contained in or derived from this unpublished research may be used by others in their writing with due acknowledgement.

2. IIUM or its library will have the right to make and transmit copies (print or electronic) for institutional and academic purposes.

3. The IIUM library will have the right to make, store in a retrieved system and supply copies of this unpublished research if requested by other universities and research libraries.

By signing this form, I acknowledged that I have read and understand the IIUM Intellectual Property Right and Commercialization policy.

Affirmed by Nu'man bin Badrud'din

……..…………………….. Signature

18 / 7 / 2021
……………………….. Date

All of my hard work is dedicated to my beloved wife, parents, family and friends for all their kindness, love and continuous support <3.

Thank you, Nur Hazwani binti Mohamad Jurimi.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Asst. Prof. Dr. Mohd. Adli bin Md. Ali and my co-supervisor Asst. Pro. Dr. Mohd. Hirzie bin Mohd. Rodzhan from the Department of Physics and CTS respectively for their continuous support and supervision during my master study and research. Their patience, motivation, enthusiasm, and immense knowledge had helped me to complete my task. Their skilled advice and constructive criticism had always kept me on toes and inspired me to thrive for excellence and follow in their footsteps.

I would like to give utmost special thanks to my dear wife, Nur Hazwani binti Mohamad Jurimi for her time and care, untiring support and encouragement throughout my journey. Truly, her support carries me throughout this journey.

Certainly, my earnest gratitude I record to my parents, Badrud'din bin Abd. Razak and Nadziroh binti Jamil, and my in-law parents Mohamad Jurimi bin Abd Hamid and Badariah binti Othman. Not forgotten my other family members who have always been generous in aspects of financial affairs and moral supports. I admit how grateful words are beyond compares to what they have sacrificed for me to achieve this stage of life.

I extend my sincere thanks to my fellow Master classmates namely Faiz, Azim, Awang, Hafidzul, Fatihah, and Sakinah for their companionship and advice. Many thanks also to the PG room members, the examiners as well as the staff officers involved along the process, who had offered their kindness and assistance in any forms, directly or indirectly throughout my research journey.

Definitely, for all their impeccable guidance and countless supports, now and then, which makes me keep pushing my limits. Indeed, they such an inspiration to me and. May God reward them with earth-full of deeds for their kindness and grant us all with good health and happiness and most expectantly the ultimate Jannah in the hereafter.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

2-D	2-dimension
3-D	3-dimension
ALICE	A Large Ion Collider Experiment
ATLAS	A Toroidal LHC ApparatuS
BSM	Beyond standard model
CERN	European Council for Nuclear Research
CMS	Compact Muon Solenoid
DDS	Distance and density structures
ESOM	Emergent Self-Organizing Map
GB	Gigabyte
GPU	Graphic processing unit
HEP	High energy physics
IDE	Integrated development environment
kNN	k-nearest neighbour
LDA	Linear discriminant analysis
LHC	Large hadron collider
LHCb	Large hadron collider beauty
Mag-grad	Magnitude gradient
ML	Machine learning
PBC	Projection-based clustering
PCA	Principal component analysis
RAID	Redundant Array of Independent Disks Mode
RAM	Random access memory
SM	Standard model of particle physics
SOM	Self-organizing map
SUSY	Supersymmetry
UCI	University of California, Irvine
U-matrix	Unified distance matrix

CHAPTER ONE

INTRODUCTION

1.1 RESEARCH BACKGROUND

Low dimensional data is constructed by samples with two or three features allowing 2-D or 3-D graph to be created. Meanwhile, high dimensional data has more than three features for each of the samples resulting in more dimensionality to exist. As the expansion of today's computing technologies are at a rapid phase through every field, the production of massive amount of data is inevitable. Hence, the analysis and interpolation of high dimensional data is becoming a crucial part in understanding the data into useful information.

A common method for analysis is to interpolate the data into graphs as a visualisation tool for further examining the traits and characteristics of the data to gain more knowledge. Nowadays, machine learning has becoming a vital part in exploratory data analysis because of its superb performance in high computing and has been implemented in many areas of research and application. One of the areas which heavily utilise machine learning is particle physics, also known as high energy physics (HEP) which mainly focuses around the study of particles as the building block of the universe.

Colliding particles via particle accelerator to study new events is a part of continuous attempt to keep exploring the limits of the standard model of particle physics (SM) into the beyond standard model (BSM) work frame. A visualizing Monte-Carlo simulation of a BSM event of supersymmetry (SUSY) by Baldi, Sadowski and Whiteson (2014) is explored in this thesis. To gain insights of the

SUSY dataset, machine learning was utilized to cluster the dataset using Self-Organizing Map (SOM). Since the BSM event is exotic from normal SM event, the SOM clustering was used for anomaly detection in which the SUSY particle is the anomaly. To verify the SUSY cluster existence in the dataset, the SOM-clustered data was then graphically represented. While U-matrix is the default visualization method for SOM, two other methods – Principal Component Analysis (PCA) and spectral graph – were attempted. Successful display of cluster existence suggests the potential of the algorithm on differentiating between SM and BSM event for further use on HEP experimentation dataset.

1.2 PROBLEM STATEMENTS

Low dimensional dataset with low number of features could just be visualized in a 2-D or 3-D graph. Meanwhile, high dimensional dataset which contains higher number of features is more challenging to be visualised. In high energy physics, the data produced by simulations and experimentations are often high dimensional. To gain further insights, the data would need to be processed further by clustering. The challenge then is to visualise the high dimensional clustered data in order to verify that the algorithm can show if there is any cluster exist in the dataset.

1.3 SIGNIFICANCE OF STUDY

This research attempts several visualization methods of Supersymmetry (SUSY) dataset that has been clustered using Self-Organizing Map (SOM) method. The visualization methods could enable us to verify the cluster existence in the clustered dataset. One of the significances of this study is a showcase of visualization of HEP dataset with different interpretations. Meanwhile, the verification of cluster existence via visualization of HEP data could provide new method for detecting abnormal detection in the dataset that hints to the existence of Beyond Standard Model (BSM) particles. Moreover, if the methods were to be optimized and expanded further perhaps it could be possible to utilize it on collision experimentation dataset.

1.4 RESEARCH OBJECTIVES

The research focuses on visualizing high dimensional HEP data and verifying cluster existence within the dataset. Therefore, the objectives of this research are:

1. To develop SOM’s U-matrix, principal component analysis (PCA) and spectral graph algorithms for visualization.
2. To visualize the SUSY dataset with different signal to background ratio using the developed algorithm.
3. To examine the capability of the developed algorithm to validate the existence of BSM event in the SUSY dataset.

1.5 RESEARCH HYPOTHESIS

Visualization of the clustered dataset using U-matrix would yield distinct graph as it is the default representation tool for SOM. Principal component analysis (PCA) graph of the dataset would be displayed in 2-D and 3-D graph because it is a dimensionality reduction algorithm. Meanwhile, spectral graph connects all the nodes together to display a network between nodes. All three methods take different approaches in observing and verifying cluster existence.

CHAPTER TWO

LITERATURE REVIEW

2.1 MACHINE LEARNING IN HIGH ENERGY PHYSICS

High energy physics (HEP) deals with tremendous amount of data, such as the Large Hadron Collider (LHC) which consumes about 40 Petabytes of storage pool considering the average file size of 200 Megabytes per file that is also replicated using RAID-1 configuration (Peter & Janyst, 2011). Approximately 1 billion times of particle collisions happens in the LHC which generates about one petabyte of collision data per second. Due to the amount of the properties of the events and particles, the data produced by HEP simulations and experimentations are routinely in high dimension.

Despite the massive data, HEP researches had significantly embraced the exposure of current technology of higher computing power and machine learning methods that pushes the boundary of previous computing limitations. Machine learning does not only influence the growth of particle physics research, but according to Albertsson et al. (2018) it is already the state-of-the-art in HEP applications such as in particle and event identification, jet pile up suppression and energy estimation. Readers interested in development areas for ML and its promising future in HEP can read the community white paper from Albertsson et al. (2018).

Therefore, implementing ML in HEP research nowadays is common as it has already become an essential tool of research and applications. In general, there are three types of ML algorithms: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning requires labelled training data for the algorithm to infer functions while unsupervised learning would only require unlabelled training data. While supervised and unsupervised learning are made to work with data samples, reinforcement learning is suited towards learning the environment as a whole. Some examples of the learning algorithms being used in HEP research are such as classification (Guest et al., 2016; Methodiev, Nachman & Thaler, 2017; Dery, Nachman, Rubbo & Schwartzman, 2017), clustering (Dokshitzer, Leder, Moretti & Webber, 1997; Dorfan, 1981), and deep learning (Guest et al., 2016; Baldi, Bauer, Eng, Sadowski & Whiteson, 2016).

Clustering is an unsupervised machine learning (ML) algorithm that gathers similar data in the dataset as a group, thus providing the user insights about the data to be interpreted and analysed. An example of cluster analysis is for anomaly detection which is achieved by training the algorithm to cluster data hence enabling the user to explore and find hidden groups, patterns or outliers in the dataset. Since the study of BSM pivots around finding new particles, HEP researches are no stranger to utilizing the anomaly detection algorithm as it is also capable of performing on high dimensional data. Examples of machine learning used for anomaly detection in HEP are one-class support vector machine (Muandet & Scholkopf, 2013), semi-supervised anomaly detection (Vatanen et al., 2012), and classifier for resonant new physics (Md Ali, Badrud'din, Abdullah, & Kemi, 2020; Collins, Howe & Nachman, 2018).

2.2 VISUALIZATION OF HIGH DIMENSIONAL DATA

The term *curse of dimensionality* was created by Bellman (1966) describing the difficulty of a problem increases very rapidly when the number of variables (dimensions) increases. This curse not only persists when solving high dimensional problem but also clustering and visualizing high dimensional data. Visualization techniques for low-dimensional spaces of typical 2-D and 3-D such as projective visualizations and parallel coordinates are ineffective against high dimensional data (Strehl & Ghosh, 2003) which means it would require other methods of visualization that are viable for high dimensional space. Further reading on this topic should include a broad survey by Liu, Maljovec, Wang, Bremer and Pascucci (2016) exploring the advancement of high dimensional data visualization that had been made within more than a decade of multitude of research works.

To overcome the curse, visualization of high dimensional data could be done in several ways. One of the techniques is dimensional reduction. Such techniques have been applied by Tang, Liu, Zhang and Mei (2016) which lays out the graph on low-dimensional space from the construction accurate approximation of nearest k-nearest neighbours (kNN) from the data. Strehl and Ghosh (2003) used relationship-based approach in visualizing similarity matrix in two dimensions from graph-partitioning-based clustered high dimensional data. Sanguinetti (2008) reduces and visualized the dimension of clustered dataset using novel probabilistic latent variable model, principal component analysis (PCA) and linear discriminant analysis (LDA).

Since the research is focused on cluster existence, one of the method suitable for this aim was the unified distance matrix (U-matrix) which can recognize cluster structures and outliers by topologically distance-mapping the input data in the data space (Ultsch, 2003). The U-matrix is the standard visualization tool for the input data distance structures of Self-organizing Map (SOM).

2.3 SELF-ORGANIZING MAP CLUSTERING

The Self-organizing system was created by Kohonen (1981) which were then evolved into the Self-Organizing Map (SOM) today. SOM is an unsupervised machine learning algorithm that projects the manifold of a high dimensional data into a low-dimensional 2-D grid (Kohonen & Somervuo, 2002). In cluster analysis, SOM network is more accurate and robust than hierarchical clustering methods in dealing with messy empirical data (Mangiameli, Chen & West, 1996). This is supported by Ultsch and Lötsch (2017) in their cluster identification in high dimensional data using machine learning which mentioned that cluster structure analysis applied using their version of SOM is unbiased and viable in contrast to using established classical hierarchical clustering algorithms which are more prone to error when identifying true clusters in the data.

Beale and Jackson (1990) described the training of the Kohonen SOM algorithm happens by first initializing the weights from the number of inputs to the nodes while also initializing the radius of the neighbourhood between the nodes. After presented with new input, the algorithm computes the distance to all nodes and for each node it selects the output with minimum distance to update the node's weight together with all other nodes in its neighbourhood. This step is then repeated for all nodes that are available, becoming self-organized by only mapping each node's distance to one another to form a tabular centroid data.

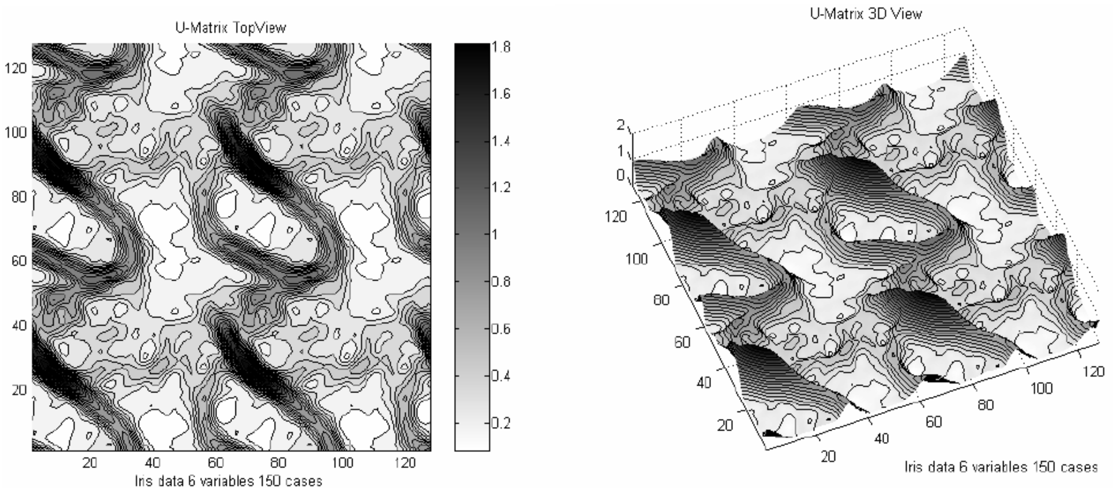

Figure 2.1 The U-matrix display for Iris data by Ultsch (2003) ESOM

Ultsch (2003) made a modified SOM model, the Emergent SOM (ESOM) and utilized U-matrix as a cluster visualization tool for their SOM. U-matrix is able to visualize SOM with high dimensional data and provides geographical interpretation because SOM preserves the topological data of the high dimensional input to be projected onto a 2-D space. An example of interpretation for SOM U-matrix would be as shown as in Figure 2.1 from their research by using terms such as "valleys" and "mountain ranges" to describe the SOM topology properties to point out cluster centres and its boundaries. The darker area is known as the "ranges" signifying the cluster boundaries while the whiter area signifies the "valleys" as the cluster centres. The number of valleys in the SOM topographic map discloses the number of clusters in the dataset (Thrun & Ultsch, 2020a).

The advantages of SOM as a clustering algorithm as described by Vesanto and Alhoniemi (2000) are:

> "First, the original data set is represented using a smaller set of prototype vectors, which allows efficient use of clustering algorithms to divide the prototypes into groups. The reduction of the computational cost is especially important for hierarchical algorithms allowing clusters

of arbitrary size and shape. Second, the 2-D grid allows rough visual presentation and interpretation of the clusters."

SOM model designed to scale with HEP data size and complexity was developed by Mohd. Adli (2017) for clustering and classification of HEP events such as supersymmetry (SUSY), Higgs and dimuon datasets. From the SOM model, SUSY dataset provided the best separation of signal and background when Euclidean similarity function was used. This thesis research adapted the SOM clustering and U-matrix approaches from Mohd. Adli (2017).

SOM also has been applied throughout many other fields with high dimensional data such as genetics (Ghouila et al., 2009), engineering (Kohonen, Oja, Simula, Visa & Kangas, 1996), document collection (Rauber, Merkl & Dittenbach, 2002), oceanography (Liu & Weisbergh, 2011), and data mining (Kiang & Kumar, 2001).

2.4 PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) introduced by Hotelling (1933) is a multivariate statistical technique. Without needing supervision, PCA could analyse variance in a dataset with many variables or high dimensional data which makes PCA a popular tool for data processing and is one of the common approaches for dimensionality reduction (Ivosev, Burton & Bonner, 2008; Murphy, 2012).

Apart of being used for dimension reduction, PCA also is no stranger to being utilised in the field of particle physics especially for analysis. For example, analysis of photon discrimination simulation of photon incident on ALICE spectrometer (Jing, Zhi-Yi, Qiu-Ying & Shu-Hua, 2010), particle tracks pattern recognition (Dutta,