Faculty of Cognitive Sciences and Human Development

A COMPARISON STUDY OF DATA CLUSTERING AND VISUALISATION TECHNIQUES WITH VARIOUS DATA TYPES

Ling Chien

Bachelor of Science with Honours
(Cognitive Science)
2020

UNIVERSITI MALAYSIA SARAWAK

Grade: A

Please tick one

Final Year Project Report ☒

Masters ☐

PhD ☐

DECLARATION OF ORIGINAL WORK

This declaration is made on the 07 day of AUGUST year 2020.

Student's Declaration:

I, LING CHIEN , 61302, FACULTY OF COGNITIVE SCIENCES AND HUMAN DEVELOPMENT, hereby declare that the work entitled, A COMPARISON STUDY OF DATA CLUSTERING AND VISUALISATION TECHNIQUES WITH VARIOUS DATA TYPES is my original work. I have not copied from any other students' work or from any other sources with the exception where due reference or acknowledgement is made explicitly in the text, nor has any part of the work been written for me by another person.

07 AUGUST 2020

LING CHIEN (61302)

Supervisor's Declaration:

I, AP. DR. TEH CHEE SIONG , hereby certify that the work entitled, A COMPARISON STUDY OF DATA CLUSTERING AND VISUALISATION TECHNIQUES WITH VARIOUS DATA TYPES was prepared by the aforementioned or above mentioned student, and was submitted to the FSKPM, UNIMAS as a *partial/~~full fulfilment~~ for the conferment of BACHELOR OF SCIENCE WITH HONOURS (COGNITIVE SCIENCE), and the aforementioned work, to the best of my knowledge, is the said student's work

Received for examination by: TC Date: 07 AUGUST 2020

(AP. DR. TEH CHEE SIONG)

I declare this Project/Thesis is classified as (Please tick (√)):

☐ **CONFIDENTIAL** (Contains confidential information under the Official Secret Act 1972)*****
☐ **RESTRICTED** (Contains restricted information as specified by the organisation where research was done)*****
☒ **OPEN ACCESS**

I declare this Project/~~Thesis~~ is to be submitted to the Centre for Academic Information Services (CAIS) and uploaded into UNIMAS Institutional Repository (UNIMAS IR) (Please tick (√)):

☒ **YES**
☐ **NO**

Validation of Project/Thesis

I hereby duly affirmed with free consent and willingness declared that this said Project/Thesis shall be placed officially in the Centre for Academic Information Services with the abide interest and rights as follows:

- This Project/Thesis is the sole legal property of Universiti Malaysia Sarawak (UNIMAS).
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis for academic and research purposes only and not for other purposes.
- The Centre for Academic Information Services has the lawful right to digitize the content to be uploaded into Local Content Database.
- The Centre for Academic Information Services has the lawful right to make copies of the Project/Thesis if required for use by other parties for academic purposes or by other Higher Learning Institutes.
- No dispute or any claim shall arise from the student himself / herself neither a third party on this Project/Thesis once it becomes the sole property of UNIMAS.
- This Project/Thesis or any material, data and information related to it shall not be distributed, published or disclosed to any party by the student himself/herself without first obtaining approval from UNIMAS.

Student's signature: _____ Supervisor's signature: _____
Date: 07 AUGUST 2020 Date: 07 AUGUST 2020

Current Address:
Lot 1716, Jalan Muhibbah, 96700 Kanowit, Sarawak.

Notes: ***** If the Project/Thesis is **CONFIDENTIAL** or **RESTRICTED**, please attach together as annexure a letter from the organisation with the date of restriction indicated, and the reasons for the confidentiality and restriction.

A COMPARISON STUDY OF DATA CLUSTERING AND VISUALISATION TECHNIQUES WITH VARIOUS DATA TYPES

LING CHIEN

This project is submitted
in partial fulfilment of the requirements for a
Bachelor of Science with Honours
(Cognitive Science)

Faculty of Cognitive Sciences and Human Development
UNIVERSITI MALAYSIA SARAWAK
(2020)

The project entitled 'A Comparison Study of Data Clustering and Visualisation Techniques with Various Data Types' was prepared by Ling Chien and submitted to the Faculty of Cognitive Sciences and Human Development in partial fulfillment of the requirements for a Bachelor of Science with Honours (Cognitive Science).

Received for examination by:

TC

(AP DR. TEH CHEE SIONG)

Date:
07 August 2020

Grade
A

ACKNOWLEDGEMENTS

First, I would like to express my deep and sincere gratitude to my research supervisor, Assoc Prof Dr. Teh Chee Siong, for giving me the opportunity to do research and providing invaluable guidance throughout this research. His dynamism, vision, sincerity, and motivation have deeply inspired me. He has taught me the methodology to carry out the research and to present the research works as clearly as possible. It was a great privilege and honour to work and study under his guidance. I am extremely grateful for what he has offered me. I would also like to thank him for his friendship, empathy, and great sense of humour.

I am extremely grateful to my parents for their love, prayers, caring, and sacrifices for educating and preparing for my future. I am very much thankful to my siblings and relatives for their love, understanding, prayers, and continuing support to complete this research work. I am so blessed to have them in my life. Immeasurable appreciation for the help are extended to volunteers and another that have contributed in completing my bachelor's degree.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Clustering is used to identify the intrinsic grouping of a set of unlabelled data. It can be applied in data mining exploration and statistical data analysis. The clustering technique plays an important role in the current digital environment. As the quality and complication of data on the internet are increasing in today's rapidly evolving area, the clustering methods become the indispensable techniques to find the patterns of the data. There are many types of clustering techniques that have been developed included partitioning methods, hierarchical clustering, density-based clustering, model-based clustering, and fuzzy clustering. This study only focuses on three types of clustering techniques which are *k*-means clustering, agglomerative hierarchical clustering with the ward's linkage, complete linkage, and average linkage, and Self-Organizing Map (SOM). The clustering algorithms are written using Python language by modifying the coding obtained from the Internet. In this project, experiments on visualisation and performance analysis of selected clustering methods are conducted. Besides that, a case study is conducted by implementing the clustering technique on online product reviews. The results for the experiment on visualisation of clustering methods, it showed that various clustering techniques have their visualisation for cluster analysis. Meanwhile, the results of the predictive accuracy indicated that *k*-means clustering and self-organizing map (SOM) are the most suitable techniques for cluster analysis. Based on the results of the case study, it concluded that the accuracy in clustering the online product reviews has the relationship with the structures and amount of the sentences. The extractive text summarisation with the clustering technique can be improved and further developed to imply in the customer review system as the correction between them have been known.

Keywords: K-means Clustering, Agglomerative Hierarchical Clustering, Self-Organizing Map (SOM), Extractive Text Summarisation with Clustering Technique

ABSTRAK

Pengelompokan digunakan untuk mengenal pasti pengelompokan intrinsik dari sekumpulan data yang tidak berlabel. Ia dapat diaplikasikan dalam eksplorasi perlombongan data dan analisis data statistik. Teknik pengelompokan memainkan peranan penting dalam persekitaran digital semasa. Oleh kerana kualiti dan kerumitan data di internet meningkat di kawasan yang berkembang pesat hari ini, kaedah pengelompokan menjadi teknik yang sangat diperlukan untuk mencari corak data. Terdapat banyak jenis teknik pengelompokan yang dikembangkan termasuk metode partisi, pengelompokan hierarki, pengelompokan berdasarkan kepadatan, pengelompokan berbasis model, dan pengelompokan kabur. Kajian ini hanya memfokuskan pada tiga jenis teknik pengelompokan iaitu kluster k-berarti, pengelompokan hierarki aglomeratif dengan kaitan bangsal, hubungan lengkap, dan perkaitan rata-rata, dan Peta Organisasi Diri (SOM). Algoritma pengelompokan ditulis menggunakan bahasa Python dengan mengubah kod yang diperoleh dari Internet. Dalam projek ini, eksperimen visualisasi dan analisis prestasi kaedah pengelompokan terpilih dilakukan. Selain itu, kajian kes dilakukan dengan menerapkan teknik pengelompokan pada tinjauan produk dalam talian. Hasil dari percobaan visualisasi kaedah kluster, menunjukkan bahawa pelbagai teknik pengelompokan mempunyai visualisasi tersendiri untuk analisis kluster. Sementara itu, hasil ketepatan ramalan menunjukkan bahawa kluster k-berarti dan Peta Organisasi Diri (SOM) adalah teknik yang paling sesuai untuk analisis kluster. Berdasarkan hasil kajian kes, disimpulkan bahawa ketepatan dalam mengumpulkan ulasan produk dalam talian mempunyai hubungan dengan struktur dan jumlah ayat. Ringkasan teks ekstraktif dengan teknik pengelompokan dapat ditingkatkan dan dikembangkan lebih jauh untuk menyiratkan dalam sistem tinjauan pelanggan kerana pembetulan antara keduanya telah diketahui.

Kata kunci: Kluster K-Berarti, Pengelompokan Hierarki Aglomeratif, Peta Organisasi Diri (SOM), Ringkasan Teks Ekstraktif Dengan Teknik Pengelompokan

CHAPTER ONE

INTRODUCTION

Introduction

Machine learning refers to a branch of computer science which studies on the learning systems (Ghahramani, 2003). Machine learning is a highly interdisciplinary area that draws on the principles of cognitive science, engineering, computer science, optimization theory, statistics, and many other science and mathematics concepts. One of the machine learning technique is unsupervised learning.

Unsupervised learning is a model which performs higher-level tasks such as classification from unlabelled input data (Coates, Ng, & Lee, 2011). It focuses on learning good feature representations through pre-training multiple layers of features using an unsupervised learning algorithm. Several design parameters are selected for each of these layers including the number of features to learn, the locations where these features are calculated, and how to encode the system's inputs and outputs (Coates, Ng, & Lee, 2011).

Clustering is a significant approach in the unsupervised learning model. Clustering is known as a data mining technique (Joshi, Kaur, & Engineering, 2013). It assigns the similar data into a cluster and dissimilar data into different clusters. There are many types of clustering techniques included partitioning methods, hierarchical clustering, density-based clustering, model-based clustering, and fuzzy clustering. The clustering algorithms can be used for data compression, model construction, organizing and categorizing data (Verma, Srivastava, Chack, Diswar, Gupta & Applications, 2012).

The clustering techniques play an important role in the current digital environment (Rao, 2003). As the quantity and complication of data on the internet are increasing in today's

rapidly evolving area, hence there is a widespread consideration on the automatic text summarization (Deshpande, Lobo, & Technology, 2013). Text summarization is a method which used to organize vast number of unstructured text documents into less manageable clusters. The stages of text summarization comprise of the topic identification, clarification, and generation of summaries. Document clustering performs an important role on the tasks like indexing, sorting, automated metadata creation, population of web source hierarchical catalogues and any operation requiring the organization of the document (Popat, Emmanuel, & technologies, 2014). The information and expertise gained can be used in a broad range of applications including technology exploration, engineering design, market research, production control, and business management (Deshpande, Lobo, & Technology, 2013).

Problem Statement

The problem arises frequently of organizing big data. According to Verma et al. (2012), it found that the groupings and categorization of data is a difficult task. This is because it requires to discover an appropriate barriers of large quantities of data in an unsupervised manner. Besides that, it needs to retain high cluster cohesiveness. For achieving this goal, it attempts to maximize similarity within clusters and reduce similarity between clusters.

On the other hand, the problem of business management occurs due to the difficulty in data organization. The dedicated reviews sites are increasing on the Web nowadays. The web-based retailers always request the reviews from the customers about the products that they had bought and the related services (Hu & Liu, 2006). It leads to the number of reviews increase significantly. Hence, the users are difficult to read and analyse them. In this study, the

purpose is to investigate possible improvements of effectiveness of document clustering by finding out various clustering algorithms available.

Objectives

There are two objectives in conducting this research as listed below:

- To conduct experiments on the visualisation and classification performance of selected clustering methods; and
- To implement the clustering technique in a case study of online product reviews.

Scope of Study

Data clustering involves partitioning the data points into multiple groups. It allows the similarity within a group is greater than the similarity among groups (Hammouda & Karray, 2000). This means that the data points to be clustered must have an underlying grouping. Otherwise, clustering of data will be failed or lead to artificially introduced partitions if the data is distributed randomly. Another problem which may occur is the overlapping of data classes (Hammouda & Karray, 2000). The performance of the clustering method will be affected by the overlapping of data groups. When the number of overlap between groupings increase, the efficiency of the clustering method decrease.

Zhan, Loh, and Liu (2009) reported the clustering-summarization approach also has the similar limitations when applied to the domain of customer reviews. The number of clusters is hard to define without prior knowledge of the set of reviews. Choosing the number of clusters improperly would eventually add noisy information and reduce performance. In clustering summarization, the set of documents is divided into non-overlapping clusters and each cluster is assumed to contain a topic. Nevertheless, there is the overlapping between the

topics frequently. They are not uniformly assigned in the non-overlapping clusters of documents. Each topic has to do with the different reviews. Similarly, each review in the set will deal with several topics rather than only one. This is because typically consumers comment on various aspects of the product instead of concentrating on one viewpoint.

Significance of Study

The clustering method aims to identify the intrinsic grouping of a set of unlabelled data (Sarada & Kumar, 2013). The main task of the clustering is not only data mining exploration, but also the common technique of statistical data analysis (Deshmukh & Gulhane, 2016). The compression of data using clustering techniques with visualisation enables the users to analyse easily (Bonner, 1964). It plays a significant role in discovering related knowledge in data. However, not all datasets are applicable to all clustering algorithms. Hence, a method for comparing the results of various clustering algorithms is essential in order to determine suitable clustering algorithms which produce the best clustering solution.

By taking advantages of the rapid development of information technology, manufacturing companies are able to collect large-scale customer information to provide strategic and technical support for their product design, development, marketing, and sales initiatives (Zhan, Loh, & Liu, 2009). Clustering technique contributes in customer review system improvement. It will increase the time efficiency outcome to a greater degree (Popat et al., 2014). Besides that, it would also help distinguish the same review from different sources. This finds the patterns and trends for evaluating the customers' interactions and offers the feedback on the shortcomings and areas to provide a better service and improve customer satisfaction (Hasan, 2018).

Definitions of Terms

Machine Learning is a set of algorithms that parse data and learns from the parsed data and use those learnings to discover patterns of interest.

Artificial Neural Network (ANN) is one set of algorithms used in machine learning for modelling the data using graphs of Neurons. It refers to a massively parallel combination of a basic processing unit which can gain knowledge from environment through a learning process and store the knowledge in its connections (Haykin, 1999).

Unsupervised Learning studies how systems can learn to interpret individual input patterns in a way that reflects the statistical structure of the overall collection of input patterns (Dayan, Sahani, & Deback, 1999).

Data Mining is an evolving set of techniques for extracting valuable information and knowledge from massive volumes of data (Alnajjar, Naser, & Control, 2015).

Clustering is the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters). It is a data mining technique to group the similar data into a cluster and dissimilar data into different clusters (Verma et al., 2012).

***K*-Means Clustering** is a common method of partitioning a data set into k groups (Wagstaff, Cardie, Rogers, & Schrödl, 2001).

Self-Organizing Map (SOM) is a neural network model for high-dimensional data analysis and visualisation (Patole, Pachghare, & Kulkarni, 2010). It maps the high-dimensional input data space onto a regular two-dimensional array of neurons.

Hierarchical Clustering creates a cluster hierarchy (a tree of clusters), also known as a dendrogram. Every cluster node has its child clusters. Sibling clusters partition the points covered by their common parent (Verma et al., 2012).

Customer Review System is the system which collects multiple online customer reviews using automatic text summarization (Zhan, Loh, & Liu, 2009).

Text Summarization refers to the reduction of a text document to generate a new form which conveys the key meaning of the contained text (Deshpande, Lobo, & Technology, 2013).

CHAPTER TWO

LITERATURE REVIEW

Introduction

Machine Learning is a sub-area of artificial intelligence (AI) technology. The system of the Machine Learning has the potential to learn independently from the experience and find the solutions to problems without the specific programming requirement. It can precisely solve various problems with complexity. The precision of the system can compete with humans or better than them. Unsupervised learning is one of the machine learning technique. It mainly deals with identifying a structure or pattern in the uncategorized data collection. Unsupervised learning contains different types of clustering techniques. However, we focus on three techniques in this study which are *k*-means clustering, hierarchical clustering, and self-organizing map (SOM). Unsupervised learning is also data mining technique which can be implemented in many systems to extract information.

Artificial Neural Networks (ANNs)

Artificial Neural Network (ANN) refers to a computational model focused on the structure of biological neural networks and their functions. ANN has the structures which built up by the interconnected adaptive artificial neurons or nodes (Basheer & Hajmeer, 2000). It can perform enormous parallel data processing and the representation of information computations. ANN-based models are empirical in nature. It can provide theoretically reliable solutions to the accurate and inaccurate formulated problems and anomalies. However, the solutions can only be understood through experimental evidence and field observations. ANN is capable of learning, generalizing, associating data, and being tolerant of fault. There are

many implementation of ANN in different applications ranging from classification, multivariate data analysis, modelling, and pattern recognition.

The structures of artificial neural network (ANN) is designed based on the operation of the biological systems and the biological neuron is the main component in building up the nervous system (Basheer & Hajmeer, 2000). It adopts the biological networking technology to solve the problems with complexity. The human nervous system has billions of neurons with different lengths and types that are relevant to their body location. There are three major functional units in a biological neuron including dendrites, cell body, and axon. Figure 1 shows the schematic of biological neuron. The dendrites play the role in receiving the signals from other neurons and transmitting the signals to cell body. The cell body contains a nucleus and a plasma. The nucleus comprises of the information on heredity traits while the plasma involves the molecular equipment used to produce the material needed by the neurons. The axon then receives cellular signals and passes them across synapses to the dendrites of neighbouring neurons (Basheer & Hajmeer, 2000).

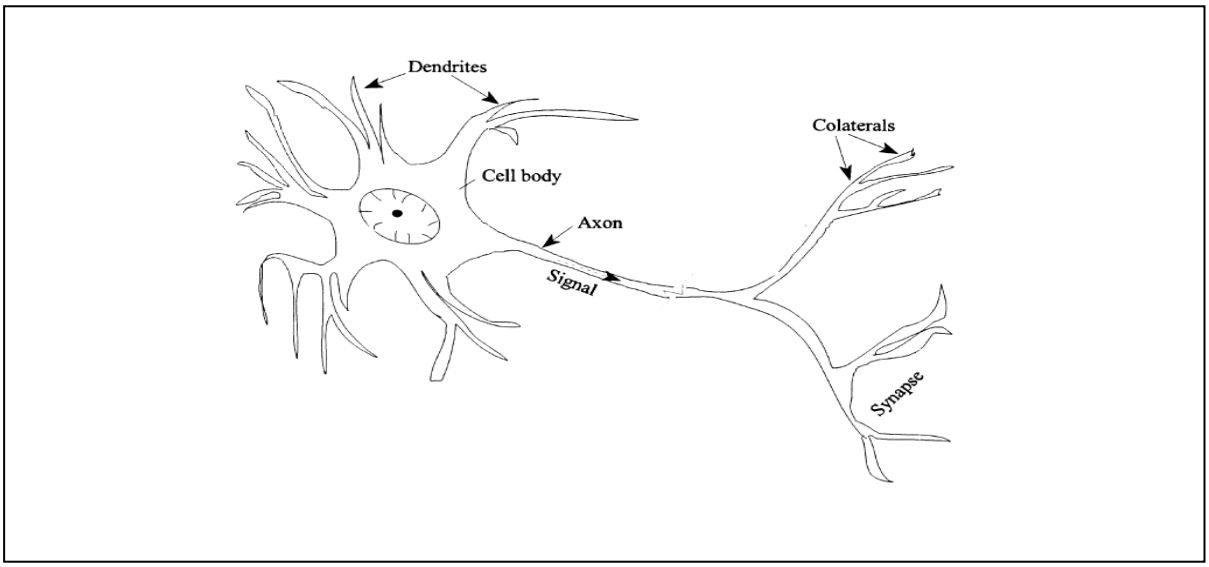

Figure 1. Schematic of Biological Neuron (Basheer & Hajmeer, 2000).

Figure 2 shows the mechanism of signal transfer between two biological neurons. An impulse in the form of an electric signal passes through the dendrites and the cell body toward the pre-synaptic membrane of the synapse (Basheer & Hajmeer, 2000). Once the signal arrives at the membrane, a neurotransmitter is released from the vesicles. The number of neurotransmitters released depends on the strength of the signal produced. The number of neurotransmitters increases when the strength of the signal produced is greater. The neurotransmitter then diffuses towards the post-synaptic membrane through the synaptic gap and passes into the dendrites of neighbouring neurons. It causes the new electrical signal to be produced. The signal generated passes through the second biological neuron and repeats the same procedures as described above. The amount of signal that passes through a receiving neuron depends on the intensity of the signal emanating from each of the feeding neurons, their synaptic strengths, and the threshold of the receiving neuron (Basheer & Hajmeer, 2000).

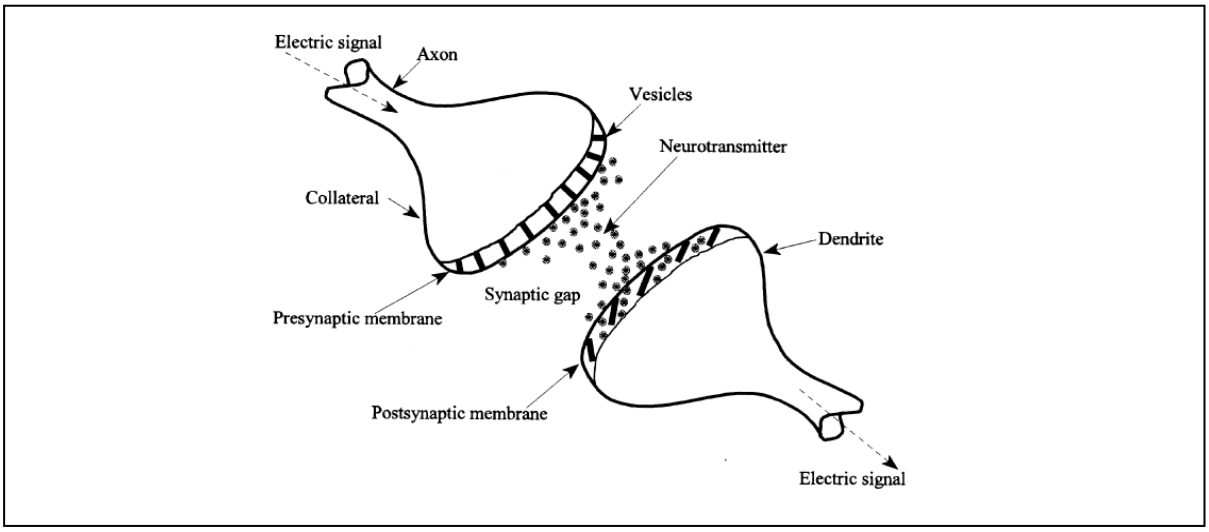

Figure 2. Mechanism of Signal Transfer between two Biological Neurons (Basheer & Hajmeer, 2000).

Figure 3 displays the signal interaction from *n* neurons and analogy to signal summing in an artificial neuron comprising the single layer perceptron. Artificial neural network

(ANN) is computational paradigms inspired by the neural structure of biological systems (Miljković, 2017). ANN uses a computational approach based on numerous artificial neurons which represent biological neurons in a simplified way. Synapses which ensure the communication between biological neurons are replaced by input weights of neurons. Learning algorithms are used to adjust the weights connection.

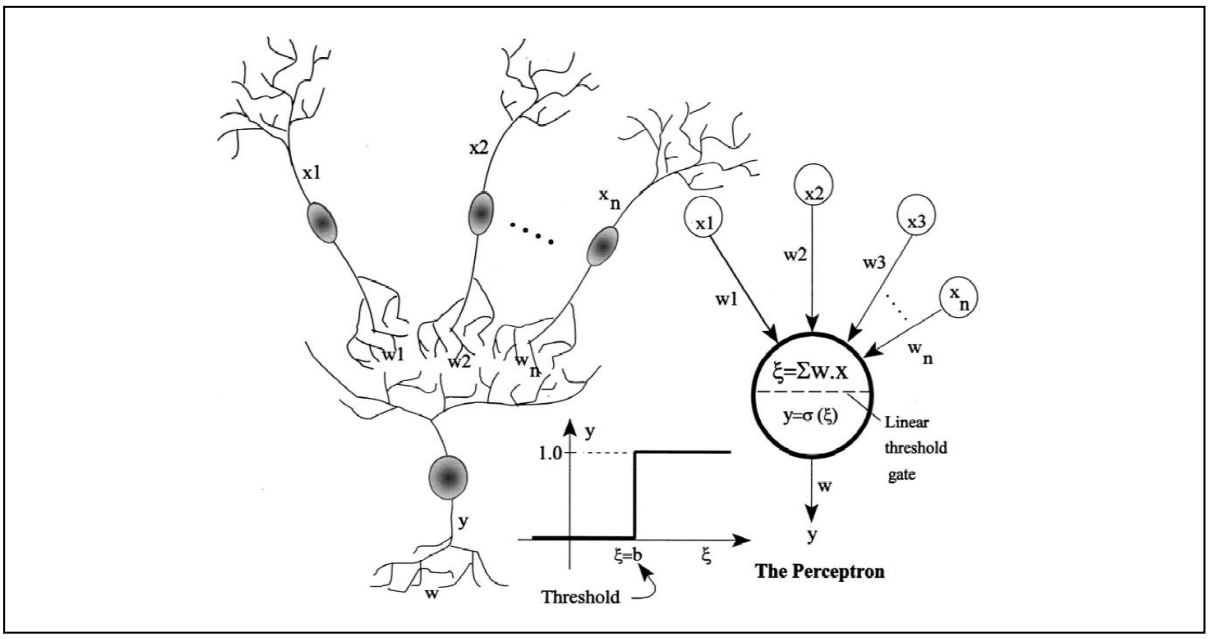

Figure 3. Signal Interaction from *n* Neurons and Analogy to Signal Summing in an Artificial Neuron comprising the Single Layer Perceptron (Basheer & Hajmeer, 2000).

Clustering

Clustering is an unsupervised learning method which used in the data analysis (Wagstaff, Cardie, Rogers, & Schrödl, 2001). Several clustering techniques have been developed and are classified based on different aspects (Joshi, Kaur, & Engineering, 2013). The categories of clustering algorithms consist of the partitioning methods, density-based methods, grid-based method, and hierarchical methods. Data collection is the first action which needs to take in the clustering process. The data collected for clustering may be numerical or categorical (Sarada & Kumar, 2013). The categorical data can be collected from

either quantitative or qualitative data where observations are explicitly derived from digits (Sarada & Kumar, 2013). The distance function between the data points of quantitative data can be described naturally by implementing the inherent geometric properties of quantitative data. The types of data used in clustering analysis consist of nominal, ordinal, binary variables, interval-scaled variables, variables of mixed types, and ratio variables.

The clustering method groups the input dataset based on similarity (Wagstaff et al., 2001). No labelled information on the dataset is provided to identify the partition of each instances. The algorithm can only access the set of characteristics that define each object. It groups the similar objects in a same cluster. However, the objects in a cluster are dissimilar to other clusters (Popat et al., 2014). Clusters can be represented in different forms including division with boundaries, spheres, probabilistic, and dendrograms (Sarada & Kumar, 2013). The efficiency of a clustering system is often measured by its ability to discover the hidden patterns of the data points. It is important to test the validity of clustering algorithms for different data types (Jin et al., 2017). This is because there are several factors may affect the result of a clustering method such as the nature of the data, algorithm selection, parameter settings, and data cleaning strategies (Jin et al., 2007). One of the parameter settings in the clustering analysis is similarity used in the process and its implementation. The similarity measure used has the significant influences in the clustering process.

According to Jin et al., (2017), there is a comparative study on the clustering techniques and its applications. The clustering techniques had been studied are Spectral Clustering, Sub-space Clustering, Adaptive *K*-means, and Self-Organizing Map. These clustering techniques are used in grouping the load profiles in order to identify their performance with various factors. Based on the results obtained, it concluded that each clustering approach has its own benefits. The result can achieve the desired properties by modifying the pre-processing and parameters of the clustering algorithm.