CLUSTERING IRRADIANCE VALUES USING UNSUPERVISED MACHINE LEARNING

YEOH YEE JUN

BACHELOR OF ELECTRICAL ENGINEERING WITH HONOURS
UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2019

CLUSTERING IRRADIANCE VALUES USING UNSUPERVISED MACHINE LEARNING

YEOH YEE JUN

A report submitted
in partial fulfilment of the requirements for the degree of
Bachelor of Electrical Engineering with Honours

Faculty of Electrical Engineering

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2019

DECLARATION

I declare that this thesis entitled “CLUSTERING IRRADIANCE VALUES USING UNSUPERVISED MACHINE LEARNING is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature :

Name : YEOH YEE JUN

Date : 24 MAY 2019

APPROVAL

I hereby declare that I have checked this report entitled “Clustering Irradiance Values Using Unsupervised Machine Learning” and in my opinion, this thesis it complies the partial fulfilment for awarding the award of the degree of Bachelor of Electrical Engineering with Honours

Signature : ....

Supervisor Name : ....

Date : ....

DEDICATIONS

To my beloved mother and father

ACKNOWLEDGEMENTS

Final year project exposes students to a world of wonder in a way that knowledge gained through our respective course might not be harnessed as it ought to be. Accountability is magnified beyond our reach and I, as a final year student, embarked on a new journey to discover the given title as a beginner. I faced a wide variety of challenges along the way and I had to seek out for help at times. Hence, I would like to express my appreciation towards a list of great people.

I wish to express my sincere gratitude to my supervisor, Mr. Kyairul Azmi Bin Baharin for exposing me to the unknown. I am grateful for his undying patience in guiding me to the right path. He showed me the right way to think in an organized manner and handle things differently. He was willing to share his experience and solution to the problems I faced without hesitation. He gave me an opportunity to go through this journey with a free will to discover the job I was held accountable for. His grit and wisdom taught me great dedication and determination. His teaching is engraved in my mindset and I shall implement them in my future undertakings.

It goes without saying that constructive comment is a huge factor in this process. I would like to thank my panel, Ms. Nurliyana Binti Baharin and Ir. Dr. Aminudin Bin Aman for commenting on my work, as well as giving suggestion for future improvement.

Finally, I would like to extend my appreciation to my friends who rendered their help directly and indirectly throughout my final year project.

ABSTRACT

Clustering is an unsupervised machine learning that works by splitting a large dataset into multiple distinctive groups. As a fast developing renewable anergy, output of photovoltaic is prone to fluctuation due to some factors. The purpose of clustering solar irradiance is to determine the daily pattern of irradiance and possibly group those having similar profile. Through this grouping, we can use the clusters obtained as a precursor for solar energy forecasting. The focus of this project is on both Self organizing map(SOM) and K-Means clustering. SOM utilizes plots for visualization purpose and to aid in manual classification. K-Means, on the other hand, makes use of silhouette analysis, elbow analysis and gap statistics analysis to determine the number of cluster. With the help of MATLAB software, a series of supporting detail and evidence is produced with minimal issue. In preliminary clustering, both SOM and K-Means are able to show similarity in the outcome, leading to a high confidence conclusion. In final clustering, an additional software, Weka is used alongside Matlab utilising only K-Means. The final outcome is the same as preliminary result where the optimum number of cluster is three. Irradiance profiles are plotted for categorization consisting of " clear sky", "cloudy" and "overcast".

ABSTRAK

Clustering adalah pembelajaran mesin tanpa pengawasan yang berfungsi dengan memisahkan kumpulan dataset yang besar ke dalam beberapa kumpulan tersendiri. Fotovoltaik merupakan penjana kuasa elektrik mesra alam sekitar yang sedang membangun. Jumlah kuasa elektrik yang dihasilkan tidak tetap. Tujuan pengelasan sinar matahari adalah untuk menentukan corak harian sinaran dan kebarangkaliannya untuk mewujudkan kumpulan yang mempunyai profil yang sama. Melalui kumpulan ini, kita boleh menggunakan kluster yang diperoleh sebagai pendahulu bagi ramalan tenaga solar. Tumpuan projek ini adalah Self Organizing (SOM) dan K-Means. SOM menggunakan plot untuk tujuan visualisasi dan untuk membantu klasifikasi secara manual. Manakala K-Means menggunakan teknik-teknik pengesahan tertentu untuk menentukan bilangan kelompok. Dengan bantuan perisian MATLAB, keputusan terperinci mampu dihasilkan tanpa isu. Pada tahap awal, kedua-dua kaedah ini dapat menunjukkan hasil yang seiras. Oleh itu, kesimpulan dapat ditentukan dengan keyakinan yang tinggi. Pada tahap terakhir, satu perisian tambahan yang bernama Weka telah diguna di samping perisian asal. Kedua-dua perisian ini dapat menunjukkan keputusan yang sama seperti pada tahap awal. Misi menentukan kategori setiap kluster dapat dijalankan dengan label “langit yang cerah”, “mendung sederhana” dan “mendung penuh”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

UTeM	-	Universiti Teknikal Malatsia Melaka
FKE	-	Faculti Kejuruteraan Elektrik
SOM	-	Self Organizing Map
KMOR	-	K-Means Outlier Removal
VRKM	-	Variance Reduced K-Means
ALO	-	Ant Lion Optimization
LM	-	Last Leap
LML	-	Last Major Leap
PCA	-	Principle Component Analysis
DSOM	-	Deep Self Organizing Map
HRSOM	-	High Resolution Self Organizing Map
LRSOM	-	Low Resolution Self Organizing Map
U-AHC	-	Agglomerative Hierarchical Clustering
MH	-	Metropolis Hastings
Qe	-	Quantization Error
Te	-	Topographic Error
U	-	Neuron Utilization
i	-	Assignment of cluster
M_i	-	New centroid
$s(i)$	-	Silhouette value
$\text{SSE}_{\text{total}}$	-	Total sum of squared error
D_r	-	Sum of pairwise distance
W_k	-	Within dispersion measure
K	-	Number of cluster
$\text{Gap}(k)$	-	Gap between reference and observation data
s_{k+1}	-	One standard error

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Photovoltaic(PV) technology is a growing industry in the past 10 years with promising rate. This trend sees large scale PV system getting integrated into the grid. However, the grid experiences more fluctuations as more PV is integrated into the grid. This phenomenon is caused by the variability of the PV output while the variability itself comes from natural causes and is uncontrollable and inevitable. Sudden changes in weather can cause immediate drop in the output. Hence, if the grid operator is not well prepared, stability problems might happen inside the grid. To address the issue, s solution through forecasting is implemented. With accurate forecasting, we can properly anticipate the changes before the actual occurrence. In order to obtain accurate forecasting, we need to classify the profile of irradiance on any day. If the classification can be done properly, then we can develop forecasting model for each weather profile. This project investigates the ability of unsupervised machine learning to automatically cluster the daily irradiance based on weather type. The main focus of this chapter is separated into five subtopics for better illustration. Project background, motivation, problem statement, objective and scope of the project are included in this proposal. Purpose of this project will be elaborated with a series of supporting details.

1.2 Project Background

Efficiency can be dubbed as a prime mover of any industry. Even the slightest fall in efficiency can deeply impact the overall income of an industry. As significant as it is, the smallest drawback can be magnified in larger industries as compared to smaller industries. Hence, the role of data scientist is highly anticipated in recent years. Given the opportunity to take up the role as a beginner in machine learning, research in the relevant field is extremely important in foundation building.

Deep comprehension of the project title plays a pivotal role in achieving the objectives. The origin, working principles, factors relating to the problem are put into account to foster understanding in depth. Photovoltaics is the direct conversion of energy obtained from the sun radiation into electricity. This takes advantage of the photoelectric effect that can be induced in some materials, causing them to release electrons when absorbing photons. Current is then induced when free electrons are captured.

Edmund Bequerel was the first person credited for the photoelectric effect in 1839, when he discovered that only specific materials would produce small amounts of current under the exposure of light. About 60 year later in 1905, Albert Einstein won a Nobel prize in physics after his breakthrough in accurately describing the fundamentals of photoelectric effect. The first photovoltaic module was created by Bell Laboratories in 1954. The module was costly at first and in the 1960s, the space industry began to make the first serious use of the technology to provide power aboard spacecraft. Space industry's involvement accelerated the technology advancement and thus, its reliability was established, and the cost began to drop. However, this technology was truly recognized in the 1970s due to energy crisis, as a source of power for non-space applications.

Figure 1.1 shown below portrays the basic operation of a solar cell. Solar cells are constructed from the same semiconductor materials, typically silicon. In solar cells, a thin semiconductor wafer is specifically treated to form an electric field, with positive and negative polarity opposite one another. Electrons are able to escape the atoms in the semiconductor material when photons reach the solar cell. If electrical conductors are connected to the positive and negative sides, completing an electrical circuit, electric current is produced due to the electron movement.

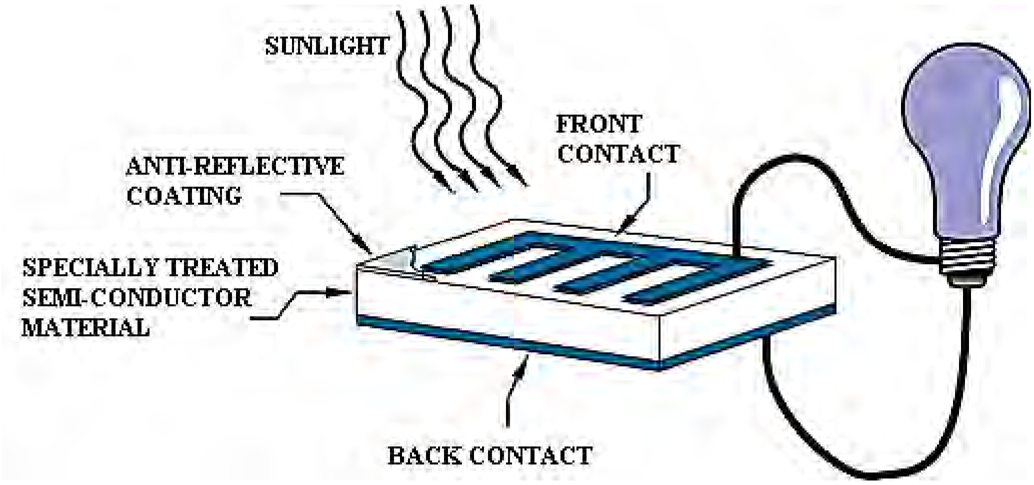

Figure 1.1: Basic operation of a photovoltaic cell

Placing a number of solar cells electrically connected to each other in a support structure or frame is called a photovoltaic module and the current induced is directly proportional to the amount of light hitting the module [1]. An illustrative diagram of photovoltaic module is shown in Figure 1.2 below.

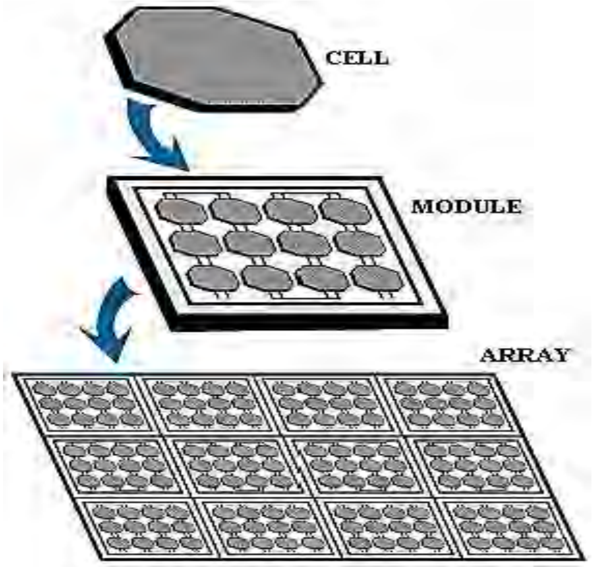

Figure 1.2: Photovoltaic module

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Clustering is important in data analysis. A good clustering algorithm is able to identity clusters irrespective of their shapes. Clustering is also known to extract information from a large data set and transform it into an understandable form for further use. The simplest stages involved in clustering algorithm are shown in Figure 1.3 [2].

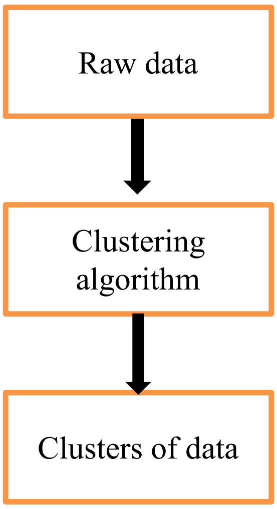

Figure 1.3: Basic clustering stage

1.3 Motivation

Photovoltaic has been around for as long as when technology advancement took place. It had a fun filled beginning that took the whole world by storm. The focus is always on the efficiency in an endeavor to powering the entire nation. Today, photovoltaics are 20% efficient and combined cycle power plants are 60% efficient. Table 1.1 shows a series of power plant type along with their efficiency [3,4,5].

Table 1.1: Efficiency of power plant type

Type	**Efficiency(%)**
Coal fired power plant	32 – 42
Natural gas fired power plant	32 – 38
Hydro turbine	85 – 90
Wind turbine	30 – 45
Solar thermal system	20
Photovoltaic system	15
Geo thermal system	35
Nuclear plant	0.27
Diesel engine	35 – 42

As we all know, photovoltaic is a better option in term of environmental friendliness. Any effort that could ease the solar related work with good accuracy is desired. Thus, having an opportunity to bring about enhancement to the solar process would ignite a sense of responsibility and willpower in me to kick start a meaningful project. Deep understanding in the relationship between solar irradiance and other variables is crucial for critical thinking in the attempt of acquiring best solution. Making full use of current software to help ease the whole lengthy process is ideal. Clustering is a known method in unsupervised machine learning and is widely used in many areas in data analysis and data mining applications. It is the task of grouping a set of objects so that objects in the same group are more similar to each other than to those in other groups. To begin with, I would use Matlab software and Weka software to discover the hidden patterns within the datasets from FKE's Solar Lab.

1.4 Problem Statement

In order to provide accurate forecasting, we need to have sufficient dataset for each weather type. Typical classification is done by separating the data into three; clear sky, cloudy and overcast. However, manual classification can be a hassle. It is especially tedious if the amount of dataset high. In each of 'big data', we already have the computational ability to process huge amount of information in a relatively short time. In term of energy forecasting, it is possible to use machine learning to do the classification. However, a lot of questions remain unanswered such as the accuracy of machine learning, data size required for machine learning to perform sufficiently and which model is the best to be used. Thus, an effort to discover the grouping, hidden patterns in solar irradiance values dataset provided by FKE's Solar Lab would help to answer the questions. A proper procedure using two known high accuracy methods, Self Organizing Map and K-Means are deployed with Matlab. At the same time, K-Means is also used in Weka software for comparison purposes.

1.5 Objectives

i) To find out the hidden pattern and strength of relationship between meteorological variables from the database in FKE’s Solar Lab.

ii) To evaluate the natural clustering of daily irradiance profiles using selected unsupervised machine learning method.

iii) To validate the results with manual classification.

1.6 Scope

a) This project utilizes Matlab R2016B software and Weka software.

b) The focus is on unsupervised machine learning, clustering.

c) The results and findings are solely based on dataset from FKE’s Solar Lab.

d) Only 9 a.m. to 7 p.m. dataset is used to avoid insignificant result.

e) One month data and three years data are used in preliminary clustering anf final clustering respectively.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter focuses on the variation of clustering method due to modification of the-state-of-the-art. Theories, benefits, comparisons and researchers' discovery will be discussed in this chapter to narrow down the options available. A litany of relevant applications dealing with real life problems are reviewed for idea expansion and to set a baseline for this project.

2.2 Machine Learning

Machine learning is a part of artificial intelligence (AI) that serves to equip systems the ability to automatically learn and improvise from past experience without being explicitly programmed. This implies the absence of specific algorithms that run specific tasks. The algorithm is better represented as shapeless with high conformity that adjusts or adapts itself in accordance to the input. The input can be in any form, ranging from numeric, shape, color, image and whichever data that have the ability to show variations [6].

The learning process is dependent on the type of algorithm. Each application requires different optimization for maximum accuracy. This process involves capturing patterns in data and learn from previous computations to produce reliable, repeatable decisions and results, thereby, producing precise predictions in the future based on the previous data [7]. The frequentative aspect of machine learning is imperative as the computed or constructed models are exposed to new input. The model is capable of adapting independently. The ultimate mission is for the computer to learn automatically without human intervention or assistance and adjust actions accordingly. Machine learning algorithms are often categorized as supervised or unsupervised.

2.2.1 Supervised Machine Learning

Supervised machine learning algorithms works by learning the pattern in the input dataset and apply the acquired model to predict the future event. Supervised learning has a prerequisite that the dataset used to train the algorithm must link to the correct answer. An inference is produced from large input dataset and is used to compare with another dataset to predict possible outcome. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly. Supervised learning usually consists of regression and classification [8]. The techniques of Supervised Machine Learning algorithms can be further split into the followings in Figure 2.1 [9].