Universiti Teknologi MARA

Web Based Clustering Tool Using K-MEAN++ Algorithm

Muhammad Nur Syazwanie Bin Aznan

2016331481

Thesis submitted in fulfilment of the requirements for Bachelor of Computer Science (Hons.)

Faculty of Computer and Mathematical Sciences

January 2019

ACKNOWLEDGEMENT

Alhamdulillah praises and thanks to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. Firstly, my special thanks goes to my supervisor, Dr. Ali bin Seman for helping me in order to complete my final year project. Special appreciation also goes to my beloved parents, who always support me from behind which have become my strength to keep moving forward. Last but not least, I would like to give my gratitude to my dearest friend because they have given me every help that I need without asking for anything.

ABSTRACT

Cluster analysis is one of the data mining task that are widely used in many area to extracting, grouping data with similar attribute in order to uncover the hidden pattern and meaning in the data. Therefore, many mathematicians and programmers have discovered many techniques and developed tools to help society to do their researches. But until today, there are only a few tools that provide the web based platform as their tools for example MATLAB MWS and Clustvis. Even though, there is a tool for clustering, some of this tool required an expert knowledge in clustering in order to understand the results. Which is why this project objective is to develop a web based clustering tool using K-MEAN++ algorithm. This project will use the rapid application development (RAD) methodology since this are the most suitable method for developing the system. The first phase in rapid application development is requirement and planning, this is where the problem statement, objective, scope, significance of this project are defined. The next phase is design, flowchart and use case diagram of this project will be design and follow accordingly. The other phases are construction where this project engine are developed using the JAVASCRIPT language and HTML for user interfaces. For testing, this project used a dummy data created with forty instances and ten attributes and Iris data contain of one hundred and fifty instances and 4 attributes. Although the tool has some limitation, but this tool are successfully tested with this two data and manage to calculate and shows the cluster with silhouette score. To sum up, these projects successfully manage to achieve the objective which is developing the web based clustering tool and all functionality are working properly.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This chapter provides the background and rationale for the study. It also gives details of the significance of web-based clustering tool, the issues and problems that led to this research

1.1 Background of Study

Cluster analysis or clustering is one of the techniques uses to group the data into different groups which the data in each group share some common trait. It is one of the common technique uses for statistical data analysis which can be uses to cluster largest data sets to identify the patterns in data. According to Berry (2004), clustering can be explained as the division of a heterogeneous data groups into homogeneous subgroups which is called "cluster" and it is a very common method in data mining in which a heterogeneous data group can be divided into significant subgroups inside. Besides, clustering analysis is an unsupervised learning method where we used for exploratory data analysis to find the hidden pattern or grouping in data.

Clustering analysis is exceptionally useful nowadays to help in grouping a set of data so it can help many areas in making decision more accurate. Moreover, clustering also gives us the ability to predict the pattern of data and prediction for new data to determine which groups are more suitable for the data. For example, UiTM students in Faculty of FSKM consist of many programs. After implementing clustering based on student's behaviour, we are able to predict for the new students who are going to choose the programme to make decision easier.

Next, the clustering analysis is widely used in many fields as it can be implement for both categorical and numerical type of data. Medical use lots of cluster analysis to handle their problems with patients and medicines. One of it