Universiti Teknologi MARA

WEB-BASED CLUSTERING TOOL USING FUZZY K-MEAN ALGORITHM

AHMAD ZULADZLAN BIN ZULKFLY

Thesis submitted in fulfillment 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 blessing, I was able to finish this research within the time duration given. Firstly, my special thanks goes to my supervisor, Dr Ali Seman for her guidance for the past semester.

Special appreciation also goes to my beloved parents Zulkfly bin Azahab and

for their endless support. I also like to give my gratitude to my dearest friend who has been a wonderful help to me in finishing this project

Thanks also to all the lecturers in the course of Bachelor of Science (Hons) Computer Science at UiTM Shah Alam for their patience and kind advice during the process of completing the project.

Last but not least, thanks you so much to all those who supporting me in any way during the completions of this proposal report by discussing, sharing or exchanging ideas and everyone who are directly or indirectly involved in writing this report.

Thank you so much.

ABSTRACT

Nowadays, clustering is one of the popular technique to grouping data to make them ease to interpret the result. Some researcher even used clustering to predict weather, natural disaster and many others. This prediction can be made by cluster the new data with the old data and observe where the data belongs. Even a normal people using clustering to grouping their data. But all the clustering tool available are not suitable enough for a normal people that does not have expert knowledge in this field. In order to make the normal people can cluster their data easily, this project aims is to develop a web-based clustering tool that can be used by all peoples. This project will use fuzzy k-means clustering algorithm to cluster the data because it is easy to implement and have many advantages. Moreover, silhouette method has been implemented to check the score of the clustering result. This project is developed by using Rapid Application Development methodology as it is the model that is most suitable for developing this web tools. This methodology consist of Requirement and Planning, Design, Construction, Testing and Implementation. In Requirement and Planning phase, the problem statement, objective and scope has been describe briefly. The study on related work and comparison of algorithm also has been done in this phase. In Design phase, the use case diagram, whole system flowchart and subsystem flowchart has been constructed to assist the development of this web tool. On the Construction phase, the development of the prototype has been started. All the algorithm for the engine has been developed by using Java script language. Lastly, on the Testing phase two type of data has been used to test this web tool result. Iris data set has been taken from UCO machine learning to assist in testing phase. The validity of the result for this web tool will be determined by the average score of the silhouette method. Although the web tool has some limitation which it can hold back the user experience on this web tool, but there are always room for improvement.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This chapter provide the background and purpose of this study. Additionally, it is also gives the details of the significance of web-based clustering tool, the issues and problem that led to this study.

1.1 BACKGROUND OF STUDY

Data is an information that has been translated into a form that can be understand by people. Data can be measured, collected, reported, analysed and visualized. Data can be divided into two types; Qualitative and quantitative. Qualitative is a data that cannot be count and it has its own categories such as gender and colour whereas Quantitative data is a data that can be count and measure. There are some data that does not have label and hard to be interpret the meaning behind it. There are many way to change the way to interpret the data.

Clustering is one of the way to easily interpret the meaning behind the data that does not have label. Clustering refers to unsupervised learning algorithm. It is a process of grouping a set of data into a set of meaningful classes called clusters. Data that are grouped together should have similar properties or attributes. Cluster analysis or clustering is a technique to grouping a set of data and can be categorized into hierarchical and partitional clustering (Xu & Wuncsh, 2008). Hierarchical clustering is one of the method in cluster analysis. This type of clustering method is to build a hierarchy of clusters where the cluster is organized as a tree. On the other hand, partitional clustering is dividing the set of data into a non-overlapping subsets.

There are many clustering algorithms that can be categorized based on their cluster model for instance fuzzy k-mean, k-AMH, EM among others. K-mean clustering is a method of grouping the object into different group which is k cluster by finding k centroids. Each of the data belongs to the group with the nearest distance to the